



ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ
ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ
Α.Α. ΜΠΛΕΤΣΑΣ

**ΕΛΛΗΝΙΚΟ ΣΥΣΤΗΜΑ
ΑΥΤΟΜΑΤΗΣ ΕΚΦΩΝΗΣΗΣ ΚΕΙΜΕΝΟΥ
(Text-To-Speech System)**



επίβλεψη: αν. καθ. Γ. ΣΕΡΓΙΑΔΗΣ

Πρόλογος

Η επιθυμία των ανθρώπων για ομιλώντες μηχανές, έγινε στην εποχή μας πραγματικότητα. Η “έκρηξη” στην εξέλιξη των ψηφιακών τεχνολογιών, εκπληρώσε το όνειρο του W. v. Kempelen, και πολλών άλλων, για μια μηχανή που μπορεί να μιλά με ανθρώπινη φωνή. Οι μηχανές αυτές ονομάζονται “ηλεκτρονικοί υπολογιστές”.

Στην εργασία αυτή ασχοληθήκαμε με τη δημιουργία ενός Συστήματος Αυτόματης Εκφώνησης Κειμένου (Text-To-Speech System), για την Ελληνική Γλώσσα.

Η μικρή εμπειρία στον Ελληνικό χώρο σε θέματα Ψηφιακής Σύνθεσης Φωνής (Digital Speech Processing), και η ακόμα μικρότερη τεχνογνωσία σε θέματα Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing), κατέστησαν την ενασχόληση με το θέμα, ιδιαίτερα επίπονη, ενώ ταυτόχρονα αποτέλεσαν και μια πρόκληση: στα πλαίσια της εργασίας αυτής, έγινε η αξιολόγηση των σπουδαιότερων τεχνικών Σύνθεσης Φωνής και Επεξεργασίας Γλώσσας υπό το πρίσμα της άμεσης προσαρμογής τους στις ιδιαιτερότητες της Ελληνικής Γλώσσας.

Μάλιστα, η συνεργασία με πεπειραμένους επιστήμονες του Εξωτερικού αποδείχθηκε εξαιρετικά χρήσιμη. Η γνωριμία και συνεργασία με τον δημιουργό του αλγορίθμου Ψηφιακής Σύνθεσης Φωνής M.B.R.O.L.A. καθηγητή T.Dutoit του Πολυτεχνείου της Mons (Βέλγιο), πάντα μέσω Διαδικτύου, βοήθησε στην προσαρμογή του αλγορίθμου αυτού, στην Ελληνική Γλώσσα.

Στο πρώτο κεφάλαιο της εργασίας αυτής, περιγράφουμε την χρησιμότητα ενός TTS, σχολιάζουμε τις βασικές λειτουργίες του και ολοκληρώνουμε με μια σύντομη περιγραφή του Ανθρώπινου Συστήματος παραγωγής ήχων.

Στο δεύτερο κεφάλαιο, περιγράφουμε την προέλευση, την φιλοσοφία και τις δυνατότητες του αλγορίθμου σύνθεσης φωνής MBROLA που χρησιμοποιήσαμε, και εν συνεχεία, σχολιάζουμε την χρησιμοποίηση της θεωρίας Γραμμικής Πρόβλεψης (L.P.C.) στην Σύνθεση Φωνής.

Στο τρίτο κεφάλαιο περιγράφουμε στην μεθοδολογία κατασκευής των φωνητικών μονάδων που χρησιμοποιήσαμε στην σύνθεση φωνής ενώ στο τέταρτο κεφάλαιο αναλύουμε τον αλγόριθμο Φωνηματικής Μεταγραφής, όπως χρησιμοποιήθηκε στον “Αίσωπο”, το Ελληνικό TTS που προέκυψε από την εργασία.

Στο πέμπτο κεφάλαιο αναλύουμε τα μοντέλα Προσωδίας που υλοποιήσαμε, περιγράφοντας πώς ο “Αίσωπος” υπολογίζει τις διάρκειες των φθόγγων, τον ρυθμό και τις

καμπύλες μεταβολής της βασικής συχνότητας των έμφωνων φθόγγων, του προς εκφώνηση κειμένου.

Στον επίλογο, αξιολογούμε το τελικό Σύστημα, χρησιμοποιώντας εκτός των άλλων και Modified Rhyme Test.

Το “ταξίδι” ήταν δύσκολο, και η “Ιθάκη” δεν θα είχε βρεθεί χωρίς την συμβολή πολύτιμων συνεργατών.

Αφετηρία της ενασχόλησης με το αντικείμενο της σύνθεσης φωνής, υπήρξε η προσεγμένη και χρήσιμη διπλωματική εργασία του συνάδελφου και φίλου Τ. Κολίτση, σε θέματα ανάλυσης του σήματος της φωνής.

Θα ήθελα να ευχαριστήσω τον συνάδελφο Γιαμαλή Α. του οποίου την φωνή δανείστηκα γι’αυτήν την εργασία. Επίσης, θα ήθελα να ευχαριστήσω τους καθηγητές κ.κ. Παναγόπουλο και Μεταξά για τις πολύτιμες συμβουλές τους.

Οι απαραίτητες ηχογραφήσεις έγιναν στο εργαστήριο Ηλεκτρακουστικής του τμήματος και επομένως θα ήθελα να ευχαριστήσω τον υπεύθυνο καθηγητή Γ. Παπανικολάου και τον ηχολήπτη κ. Κόντο.

Η προσαρμογή του MBROLA αλγορίθμου στα Ελληνικά, έγινε με την βοήθεια του ιδιαίτερα συνεργάσιμου Τ.Dutoit και του συνεργάτη του V.Pagel, του Πολυτεχνείου της Mons.

Η υλικοτεχνική υποστήριξη από τον computer director του Media Lab (M.I.T.) Μ.Μπλέτσα, υπήρξε επίσης ιδιαίτερα σημαντική.

Στην εργασία αυτή, πολύτιμος συνεργάτης υπήρξε και η μεταπτυχιακή φοιτήτρια του τμήματος Β. Οικονομίδου, η οποία βοήθησε ιδιαίτερα με τις χρήσιμες συμβουλές της στην προγραμματιστική υλοποίηση του “Αίσωπου”.

Τέλος, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου (και τελικώς, καλό φίλο) Γ. Σεργιάδη, ο οποίος ‘αγκάλιασε’ θερμά την όλη προσπάθεια και λειτούργησε καταλυτικά με τις πολύτιμες γνώσεις του.

Άγγελος Μπλέτσας
Θεσσαλονίκη, Νοέμβριος ’98

Περιεχόμενα

Κεφάλαιο 1: Εισαγωγή	1
1.1 Εφαρμογές ενός Text-to-Speech Συστήματος	1
1.2 Βασικές λειτουργίες ενός Text-to-Speech Συστήματος	3
1.3 Ο Ανθρώπινος Μηχανισμός Παραγωγής Φωνής	5
 Κεφάλαιο 2: Αλγόριθμοι Ψηφιακής Σύνθεσης Φωνής	 10
μέρος 1: M.B.R.O.L.A.	
2.1 Γενική αρχή αλγόριθμου MBROLA	10
2.2 Προέλευση-Φιλοσοφία	10
2.3 Δυνατότητες MBROLA τεχνικής	16
μέρος 2: L.P.C.	21
2.4 LPC : Γενική αρχή - μέθοδος Ανάλυσης της Φωνή	21
2.5 Μέθοδοι Υπολογισμού Συντελεστών πρόβλεψης ακ	25
2.6 LPC και Σύνθεση Φωνής	26
 Κεφάλαιο 3: Παραγωγή Ακουστικών-Φωνητικών Μονάδων	 32
3.1 Είδος Φωνητικών Μονάδων - Γιατί Δίφωνα;	32
3.2 Συμβολισμός Φωνητικών μονάδων	34
3.2.1 Βασικές Έννοιες	34
3.2.2 Κατάταξη των φθόγγων της Ελληνικής	35
3.2.3 Το Φωνηματικό Αλφάβητο του “Αίσωπου”	37
3.2.4 Καταγραφή Φωνητικών-Ακουστικών Μονάδων	39
3.3 Εξαγωγή Διφώνων	40
 Κεφάλαιο 4: Φωνηματική Μεταγραφή	 45
4.1 Τεχνικές Φωνηματικής Μεταγραφής - Επιλογή μεθόδου για τον “Αίσωπο”	45
4.2 Δυνατότητες του Αλγορίθμου Φωνηματικής Μεταγραφής	46
4.3 Περιγραφή Αλγορίθμου Φωνηματικής Μεταγραφής	47
 Κεφάλαιο 5: Μοντέλα Προσωδίας	 54
5.1 Βασικές έννοιες	54

5.1.1	ΤΟΝΙΣΜΟΣ - Τόνος Δυναμικός, Επιτόνηση, Ποιότητα-Διάρκεια	54
5.1.2	Τα Σημεία στίξης	55
5.2	Τα Προσωδικά Μοντέλα στον “Αίσωπο”	56
5.2.1	Μοντέλα διάρκειας (Timing)	57
5.2.1.1	Υπολογισμός μέσης τιμής και τυπικής απόκλισης της διάρκειας των φωνημάτων	59
5.2.1.2	Sum of Products Μοντέλο Διάρκειας (S.o.P)	62
5.2.2	Μοντέλα Επιτόνισης (Intonation)	63
5.2.2.1	Μηχανισμός εύρεσης παύσεων	64
5.2.2.2	Linear Intonation models στον “Αίσωπο”	66
	Κεφάλαιο 6: Επίλογος	71
6.1	Η τελική δομή	71
6.2	Αξιολόγηση	73
6.3	Μελλοντικές Βελτιώσεις	76
	Παράρτημα	78
	Βιβλιογραφία	81

Κεφάλαιο 1

Εισαγωγή

Γλώσσα είναι η δυνατότητα που έχει ο Άνθρωπος να εκφράζει της σκέψεις του χρησιμοποιώντας ένα σύνολο “σημαδιών”, είτε είναι αυτά γραφικά (όπως οι Λατινικοί χαρακτήρες ή τα Αραβικά ιδεογράμματα), είτε είναι χειρονομίες (όπως η γλώσσα των βωβών ή κωφάλαλων), είτε είναι μουσικά σημάδια (όπως η γλώσσα με σφυρίγματα που χρησιμοποιούν οι βοσκοί στα Πυρηναία) είτε είναι φθόγγοι, τα ακουστικά σημάδια μιας φωνής.

Επομένως η Γλώσσα, είναι ένα διακεκριμένο χαρακτηριστικό του Ανθρώπου, καθώς μόνο αυτός μπορεί να χρησιμοποιήσει ένα τόσο αυστηρά δομημένο **σύστημα**. Η **φωνή**, είναι ένα από τα απαραίτητα συστατικά αυτού του συστήματος. Αποτελεί μάλιστα, τον αρχαιότερο και τον πιο διαδεδομένο τρόπο επικοινωνίας μεταξύ των ανθρώπων.

Δεν αποτελεί έκπληξη επομένως, το γεγονός ότι οι άνθρωποι μελέτησαν την φωνή εκτενώς και σε ορισμένες περιπτώσεις, προσπάθησαν να φτιάξουν **μηχανές** που να παράγουν φωνή με **αυτόματο τρόπο**. Χαρακτηριστική ήταν η προσπάθεια του Wolfgang von Kempelen, το 18^ο αιώνα, φωτογραφίες και σχέδια της οποίας, παρουσιάζονται στο παράρτημα.

Τα εντυπωσιακά άλματα τεχνολογικής προόδου, από την εποχή που κατασκευάστηκε το “φουσητήρι” του v.Kempelen μέχρι σήμερα, πραγματοποίησαν το “όνειρο”: υπάρχουν πλέον, μηχανές, που μπορούν να μιλούν με Ανθρώπινη Φωνή, και μάλιστα, όσο η τεχνολογία προχωρά, η συνθετική φωνή βελτιώνεται και μοιάζει περισσότερο στην ανθρώπινη. Οι μηχανές αυτές, ονομάζονται **ηλεκτρονικοί υπολογιστές**.

Στην παρούσα εργασία, ασχοληθήκαμε με την ανάπτυξη ενός Συστήματος Αυτόματης Εκφώνησης κειμένου της Ελληνικής Γλώσσας, (Text-To-Speech System, TTS) το οποίο το ονομάσαμε **“Αίσωπο”** (*Experimental Study on Prosody of Speech, E.S.o.P.o.S.-TTS*).

Η προσπάθεια αυτή, επικεντρώνεται στη ΝεοΕλληνική Γλώσσα και επιχειρεί να καλύψει ή πιο σωστά, παλεύει, με ένα μεγάλο κενό της Ελληνικής Επιστημονικής κοινότητας, στην τεχνογνωσία των T.T.S. συστημάτων.

Στο υπόλοιπο της ενότητας αυτής, θα περιγράψουμε την χρησιμότητα ενός Text-To-Speech συστήματος, θα σχολιάσουμε τις βασικές λειτουργίες που οφείλει να επιτελεί και τέλος, θα περιγράψουμε εν συντομία τον τρόπο με τον οποίο ο Άνθρωπος παράγει φωνή.

1.1 Εφαρμογές ενός Text-to-Speech Συστήματος

Βοήθημα για ανθρώπους με ειδικές ανάγκες

Ένα TTS σύστημα μπορεί να βοηθήσει εξαιρετικά ανθρώπους με *προβλήματα ομιλίας* εξαιτίας μηχανικών / νευροαισθητηριακών ανωμαλιών. Ο διάσημος αστροφυσικός **Stephen Hawking** μπορεί να μιλήσει μόνο με την βοήθεια ενός τέτοιου συστήματος.

Επιπλέον, ένα text-to-speech σύστημα οδηγεί στην πρόσβαση ψηφιακών κειμένων (ή μη ψηφιακών με την βοήθεια ενός συστήματος οπτικής αναγνώρισης χαρακτήρων - Optical Character Recognition) από **ΤΥΦΛΟΥΣ ανθρώπους ή αμβλύωπες**.

Τηλεπικοινωνιακές υπηρεσίες**ι) Συστήματα ανάκτησης πληροφορίας (Information-retrieval Systems)**

Τα TTS συστήματα καθιστούν δυνατή την πρόσβαση σε γραπτή πληροφορία μέσω τηλεφώνου. Η παραπάνω προοπτική αξίζει ιδιαίτερης προσοχής γνωρίζοντας πως το 70%, περίπου, των τηλεφωνημάτων απαιτούν μικρή αλληλεπίδραση. Η γραπτή πληροφορία μπορεί να ποικίλει από απλά μηνύματα τοπικών πολιτιστικών γεγονότων (κινηματογράφοι, θέατρα) μέχρι τεράστιες βάσεις δεδομένων που δύσκολα μπορούν να διαβαστούν και να αποθηκευτούν μαγνητοφωνημένες. Οι αναζητήσεις στα παραπάνω συστήματα ανάκτησης πληροφορίας (information - retrieval systems) μπορεί να γίνει με την βοήθεια της φωνής του ενδιαφερόμενου (μέσω αναγνώρισης φωνής) ή μέσω του τηλεφωνικού πληκτρολογίου (DTMF συστήματα).

ιι) Εξελιγμένη Μετάδοση Μηνυμάτων (Integrated Messaging)

Ιδιαίτερος υποσχόμενη εφαρμογή ενός TTS συστήματος είναι η αυτόματη ανάγνωση ηλεκτρονικών γραμμάτων (e-mail) ή τηλεομοιοτύπων (telefax) μέσω τηλεφώνου, **χωρίς την χρήση υπολογιστή ή άλλης εξειδικευμένης συσκευής!**

Προφορική Παρακολούθηση (Vocal Monitoring)

Σε ορισμένες περιπτώσεις τα προφορικά μηνύματα είναι αποτελεσματικότερα των γραπτών καθώς δεν απαιτούν οπτική προσοχή. Επομένως ένα text-to-speech σύστημα θα ήταν χρήσιμο σε συστήματα μέτρησης και ελέγχου (π.χ. σε cockpit αεροπλάνου ή σε ανακοινώσεις αεροδρομίων)

Εκπαίδευση Γλώσσας και Εφαρμοσμένη Έρευνα (Language Education, applied Research)

Ένα υψηλής ποιότητας TTS σύστημα μπορεί να αποδειχθεί ιδιαίτερος χρήσιμος στην διδασκαλία μιας γλώσσας.

Επιπλέον ο καθολικός έλεγχος του χρήστη πάνω στο σύστημα επιτρέπει την επανάληψη

γλωσσικών-φωνητικών πειραμάτων με **σταθερά** αποτελέσματα, γεγονός αδύνατον όταν αντί για ψηφιακό συνθέτη χρησιμοποιείται άνθρωπος.

Συνεπώς ένας συνθέτης φωνής αποτελεί πολύτιμο εργαστηριακό βοήθημα για έναν υπολογιστικό γλωσσολόγο ή φωνολόγο.

Με τον “Αίσωπο” για παράδειγμα, καταλήξαμε πως η μικρότερη μεταβολή συχνότητας που μπορεί να αντιληφθεί το ανθρώπινο αυτί στην περιοχή των 100 Hz, είναι περίπου 5 Hz.

Εφαρμογές Multimedia, Ομιλώντα παιχνίδια, επικοινωνία Ανθρώπου-μηχανής

1.2 Βασικές λειτουργίες ενός Text-to-Speech Συστήματος

Κάθε TTS σύστημα, δέχεται ως είσοδο κείμενο και το μετατρέπει σε φωνή. Συνεπώς, στην σχεδίαση του συστήματος γίνεται *a priori* η παραδοχή πως το κείμενο εισόδου, είναι συμβατό, τουλάχιστον, με τους γραμματικούς κανόνες του Γραπτού Λόγου της Γλώσσας.

Οι βασικές λειτουργίες ενός TTS συστήματος, το οποίο υλοποιείται με την βοήθεια ενός ηλεκτρονικού υπολογιστή, είναι τρεις:

- i) ***Φωνηματική μεταγραφή***
- ii) ***Υπολογισμός Προσωδίας***
- iii) ***Ψηφιακή Σύνθεση Φωνής***

Η φωνηματική μεταγραφή, θα ενσωματώσει το σύνολο των φωνηματικών κανόνων της γλώσσας έτσι ώστε να μετατρέψει χαρακτήρες ή/και συνδυασμούς χαρακτήρων του κειμένου εισόδου, στα κατάλληλα φωνήματα (φωνήματα = σύμβολα των φθόγγων της κάθε γλώσσας), στα κατάλληλα δηλαδή φωνητικά-ακουστικά σημάδια. Έτσι, οι μηχανισμοί φωνηματικής μεταγραφής, την λέξη <<σβέλτος>> θα πρέπει να την μετατρέψουν σε /zveltos/ και όχι σε /sveltos/.

Η αντιστοιχία γραμμάτων-φωνημάτων δεν είναι αμφιμονοσήμαντη, εξαιτίας του διαφορετικού ρυθμού με τον οποίο εξελίσσεται ο Προφορικός Λόγος σε σχέση με τον Γραπτό. Έτσι, ο μηχανισμός φωνηματικής μεταγραφής θα πρέπει να χειρίζεται σωστά περιπτώσεις όπως στην λέξη <<πιεστικός>> ➔ /πι-ε-στικός/¹ ενώ <<πιες>> ➔ /πχές/.

Επιπλέον, ο αλγόριθμος φωνηματικής μεταγραφής θα πρέπει να χειρίζεται ειδικούς χαρακτήρες, όπως οι συντομογραφίες και τα νούμερα, τα οποία προφανώς, πρέπει πρώτα να μετατραπούν σε λέξεις στην σωστή κλίση, και στην συνέχεια σε φωνή.

¹ με το αλφάβητο των συμβόλων των φωνημάτων, θα ασχοληθούμε σε επόμενα κεφάλαια

Ο μηχανισμός προσωδίας, είναι υπεύθυνος για την απόδοση διάρκειας σε κάθε φωνητική μονάδα, τον υπολογισμό των μεταβολών της βασικής συχνότητας $F0^2$ στο προς εκφώνηση κείμενο (intonation) καθώς επίσης και για την εκτίμηση των σημείων του κειμένου, όπου πρέπει να δοθεί έμφαση (stress-accentuation).

Η προσωδία ενός κειμένου, δεν είναι απαραίτητη μόνο για την “μουσικά χρωματισμένη”, όχι μονότονη ($F0=const$) εκφώνηση του κειμένου, αλλά είναι ιδιαίτερα σημαντική εξαιτίας της γλωσσολογικής πληροφορίας που μεταφέρει, καθώς κατάλληλες προσωδικοί παράμετροι μπορούν να εκφράσουν ερώτηση ή κατάφαση ή ακόμη να υποδηλώσουν την ψυχική κατάσταση του ομιλητή.

Τους μηχανισμούς υπολογισμού προσωδίας για τη σωστή εκφώνηση ενός κειμένου, ο Άνθρωπος τους αφομοιώνει υποσυνείδητα, κατά την εκμάθηση της Γλώσσας. Υπάρχει μάλιστα και ειδική επιστήμη, η Φωνολογία, που ασχολείται ακριβώς με το πώς ο Άνθρωπος, υπολογίζει προτού διαβάσει την προσωδία ενός κειμένου.

Είναι κατανοητό λοιπόν, πως η εκπαίδευση μιας μηχανής στο πώς θα υπολογίζει τις προσωδικές παραμέτρους ενός κειμένου, ενός οποιοδήποτε κειμένου, δεν είναι μια εύκολη υπόθεση. Ο υπολογισμός της Προσωδίας και η φωνηματική μεταγραφή, αποτελούν αντικείμενα της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing).

Στο τελευταίο στάδιο επεξεργασίας, επιτελείται η Ψηφιακή Σύνθεση Φωνής. Ακριβώς επειδή ο μηχανισμός παραγωγής φωνής είναι ο ίδιος για όλους τους Ανθρώπους, ενώ οι Γλώσσες διαφέρουν, οι τεχνικές Ψηφιακής Σύνθεσης Φωνής, έχουν μελετηθεί και βελτιωθεί περισσότερο, απ’ότι οι αλγόριθμοι Επεξεργασίας Φυσικής Γλώσσας.

Υπάρχουν σήμερα, Συνθέτες φωνής που χρησιμοποιούν αλυσιδωτή σύνδεση προηχογραφημένων ψηφίδων φωνής (concatenation synthesis) αλλά και παραμετρικοί συνθέτες, οι οποίοι χρησιμοποιούν ψηφιακά, χρονομεταβλητά φίλτρα με απόκριση συχνότητας κατάλληλη έτσι ώστε να δημιουργηθεί συνθετική φωνή, παραπλήσια φασματικά με την ανθρώπινη (formant synthesizers).

Επίσης, υπάρχουν Συνθέτες, που προσπαθούν να μιμηθούν την μορφή του σήματος της Ανθρώπινης φωνής, και συνθέτες που προσπαθούν να μιμηθούν τον Ανθρώπινο τρόπο παραγωγής φωνής. Στην τελευταία κατηγορία, ανήκουν οι Συνθέτες άρθρωσης (articulation synthesizers), οι οποίοι με μαθηματικά μοντέλα προσομοιώνουν τις κινήσεις όλων των αρθρωτήρων του φωνητικού σωλήνα (vocal tract), για την παραγωγή

κάθε φθόγγου.

² της συχνότητας δηλαδή με την οποία ταλαντώνονται οι φωνητικές χορδές στους έμφωνους φθόγγους

Παραδείγματα *concatenation synthesis* αποτελούν ο “Αίσωπος”, το σύστημα της AT&T κ.α. ενώ παράδειγμα *formant synthesis* αποτελεί το MITalk κ.α.

Παράδειγμα *articulation synthesis* αποτελεί η εργασία του καθηγητή P.Cook στο πανεπιστήμιο του Princeton (<http://www.cs.princeton.edu/~prc/>).

Στον “Αίσωπο” η Φυσική Επεξεργασία Γλώσσας των φράσεων <<καλημέρα σας. Τι κάνετε;>> θα έδινε τα παρακάτω, τα οποία θα τροφοδοτούσαν τον συνθέτη φωνής:

```
_ 128
k 60
a 58 0 120
l 43
i 35 100 120
m 85 100 92
e 97 100 80
r 24
a 60 100 80
s 112
a 64 100 80
s 226
_ 125
t 40
i 40 0 120 100 112
k 40
a 98 0 105 100 116
n 40 0 98 100 95
e 54 100 95
t 89
e 115 0 126 100 161
_ 195
```

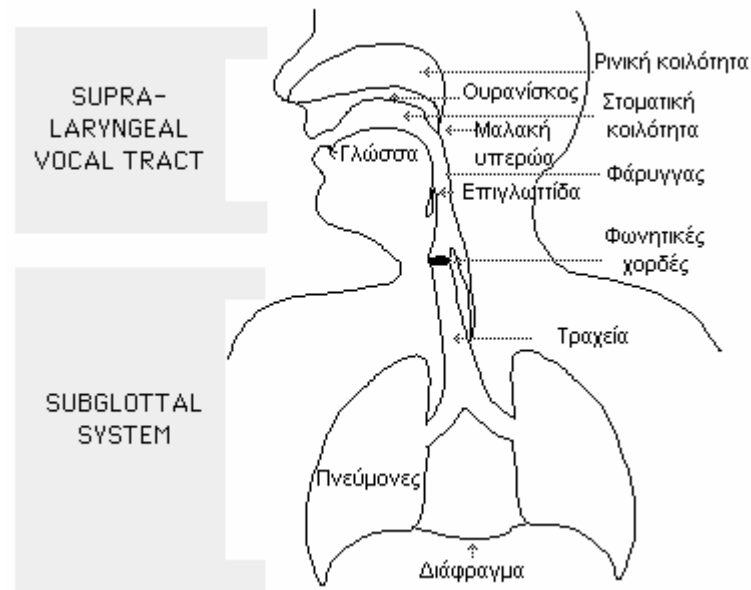
Η πρώτη στήλη περιέχει τη φωνηματική μεταγραφή, το πρώτο νούμερο σε κάθε γραμμή αποτελεί την διάρκεια (σε msec) του αντίστοιχου φωνήματος ενώ το υπόλοιπο string αποτελεί την πληροφορία του intonation π.χ η γραμμή **i 40 0 120 100 112** σημαίνει φώνημα /i/ διάρκειας 40 msec, ξεκινώντας με βασική συχνότητα 120 Hz στο 0% της διάρκειάς του, και καταλήγοντας γραμμικά στα 112 Hz, στο 100% της διάρκειας του φωνήματος /i/.

Στα επόμενα κεφάλαια, θα εξηγήσουμε την λειτουργία του Συνθέτη φωνής και θα περιγράψουμε τους μηχανισμούς Φωνηματικής Μεταγραφής και Προσωδίας στον Αίσωπο, για κάθε κείμενο.

Ακολουθεί, για λόγους πληρότητας, η συνοπτική περιγραφή του Ανθρώπινου Συστήματος παραγωγής ήχων.

1.3 Ο Ανθρώπινος Μηχανισμός Παραγωγής Φωνής

Στο παρακάτω σχήμα, παρουσιάζεται το ανθρώπινο σύστημα παραγωγής ήχου.



Υπάρχουν πολλοί τρόποι παραγωγής ήχου:

Έμφωνοι φθόγγοι (voiced phones): μια μέθοδος σχετίζεται με την χρησιμοποίηση της πίεσης αέρα, ο οποίος παρέχεται από τους πνεύμονες και θέτει τις ελαστικές φωνητικές χορδές σε ταλάντωση. Επομένως, ο λάρυγγας, στον οποίο βρίσκονται οι φωνητικές χορδές, μετατρέπει την σταθερή ροή αέρα που προέρχεται από το υπογλωττιδικό³ σύστημα (subglottal system), σε μια σειρά από παλμούς (puffs), με συνέπεια ένα σχεδόν περιοδικό (quasi-periodic) κύμα ήχου.

Άφωνοι φθόγγοι (unvoiced phones): μια δεύτερη μέθοδος, αφορά τους απεριοδικούς ήχους, οι οποίοι παράγονται από το ελεύθερο πέρασμα του αέρα μέσα από το άνοιγμα των φωνητικών χορδών (γλωττίδα), χωρίς οι τελευταίες να ανοιγοκλείνουν. Το ρεύμα αέρα, ‘διαμορφώνεται’ από τον υπερλάρυγγικό φωνητικό σωλήνα (supra-laryngeal vocal tract), λόγω στροβιλώδους ροής σε τοπικά σημεία περιορισμού (vocal tract constrictions).

Μεταβατικοί φθόγγοι (transient phones): η τρίτη μέθοδος αφορά την παραγωγή μεταβατικών ήχων, που προκαλούνται από την αιφνίδια αλλαγή του σχήματος του φωνητικού σωλήνα (vocal tract). Οι ήχοι της κατηγορίας αυτής, είναι και οι πιο δύσκολοι στη “Σύνθεση Φωνής”.

Το ανοιγοκλείσιμο των φωνητικών χορδών στην περίπτωση των έμφωνων φθόγων, προκύπτει όταν η υπογλωττιδική πίεση αέρα, είναι μεγαλύτερη από την αντίσταση των φωνητικών χορδών. Σ’αυτή την περίπτωση, οι φωνητικές χορδές ανοίγουν ώστε να απελευθερωθεί ο αέρας, και κλείνουν ξανά εξαιτίας της ελαστικότητάς τους και του φαινομένου Bernoulli. Εφόσον η υπογλωττιδική παροχή αέρα υπό πίεση συνεχίζεται, οι

³ η γλωττίδα είναι το τριγωνικό πέρασμα που σχηματίζεται κατά το άνοιγμα των φωνητικών χορδών

φωνητικές χορδές ανοιγοκλείνουν-ταλαντώνουν, παράγοντας τους σχεδόν-περιοδικούς παλμούς, με συχνότητα που ονομάζεται **βασική συχνότητα F_0** της λαρυγγικής πηγής (fundamental frequency of laryngeal source) και η οποία καθορίζει τον τόνο (pitch) με τον οποίο αντιλαμβανόμαστε τον κάθε παραγόμενο ήχο.

Κατά την διάρκεια της φυσικής ομιλίας, η F_0 μεταβάλλεται διαρκώς, προσφέροντας γλωσσολογικές πληροφορίες, καθώς η μεταβολή της βασικής συχνότητας συναρτήσει του χρόνου (intonation), σχετίζεται με την έκφραση ερώτησης ή κατάφασης, ή την εκδήλωση συναισθήματος κατά την διάρκεια της ομιλίας. Συνεπώς, η ενσωμάτωση μηχανισμών χειρισμού των μεταβολών της F_0 , είναι θεμελιώδης σε ένα Σύστημα Σύνθεσης Φωνής.

Το ρεύμα αέρα που προέρχεται από τους πνεύμονες, διαμορφωμένο από τις φωνητικές χορδές (quasi-periodical) ή όχι, εξέρχεται διαμέσου της στοματικής ή της ρινικής κοιλότητας ή και των δύο. Το σχήμα του φωνητικού σωλήνα, η διάταξη δηλ. των κινητών αρθρωτήρων του φωνητικού συστήματος είναι χαρακτηριστική για κάθε φθόγγο, και μπορούμε να πούμε πως αντιστοιχεί σε ένα χρονομεταβλητό φίλτρο-αντηχείο, με συγκεκριμένες συχνότητες συντονισμού για κάθε φθόγγο. Οι συχνότητες αυτές, ονομάζονται **συχνότητες διαμόρφωσης (formants)** και αποτελούν μια διαφορετική, μαθηματική θα λέγαμε, αναπαράσταση του κάθε φθόγγου.

Επομένως, στο μοντέλο παραγωγής φωνής που υιοθετούμε, η στοματική και η ρινική κοιλότητα, σε συνδυασμό με την θέση των κινητών αρθρωτήρων (σαγόني, γλώσσα) αντιστοιχούν σε ένα φίλτρο, του οποίου η απόκριση συχνότητας (όπως καθορίζεται από τις συχνότητες διαμόρφωσης του κάθε φθόγγου) σχηματοποιεί το φάσμα του σήματος διέγερσης, το οποίο μπορεί να είναι περιοδικό ή όχι, συντελώντας στην προφορά του κάθε ήχου. Το μοντέλο αυτό, υποθέτει την ανεξαρτησία μεταξύ των formants και του σήματος διέγερσης. Η παραδοχή αυτή είναι εν μέρει σωστή και θα συζητηθεί περαιτέρω στο κεφάλαιο Κωδικοποίησης Γραμμικής Πρόβλεψης.

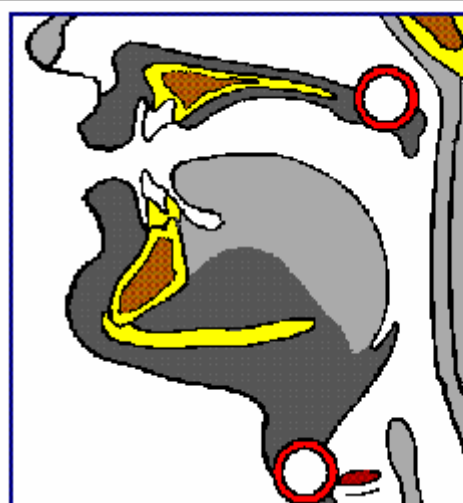
Η κατηγοριοποίηση των φθόγγων της Ελληνικής, θα γίνει στο κεφάλαιο Παραγωγής Φωνητικών Μονάδων.



Βασικές περιοχές
περιορισμού
(constriction areas)

1. κλείσιμο=κροτικά

2. στένεμα=τριβικά

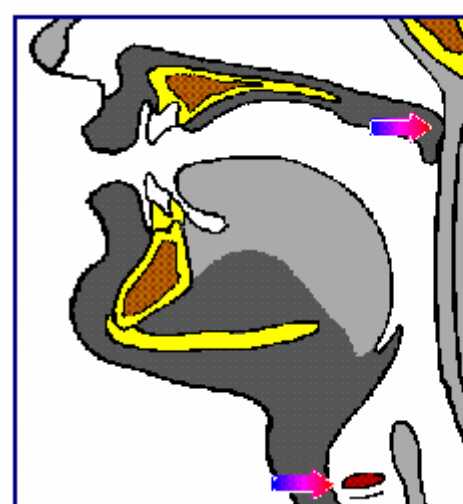


Ταλάντωση
φωνητικών
χορδών &
σύμφωνα

π.χ
ρινικό σύμφωνο /m/

τριβικό σύμφωνο /z/

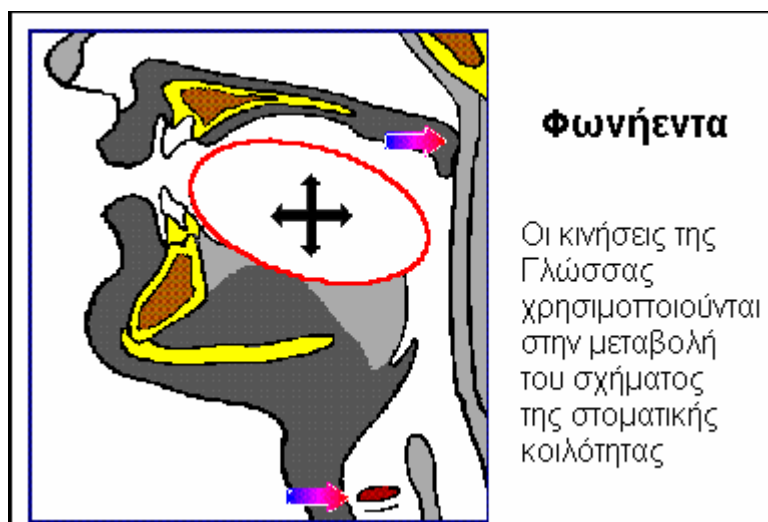
κροτικό σύμφωνο /b/



Φωνήεντα

Η μαλακή υπερώα
(velum)
σηκώνεται,
εμποδίζοντας
το ρεύμα αέρα
να εισέλθει
στην ρινική κοιλότητα.
Υπάρχουν, ωστόσο,
και ρινικά φωνήεντα.

οι φωνητικές χορδές
ταλαντώνονται



Κεφάλαιο 2

Ψηφιακή Σύνθεση Φωνής

μέρος πρώτο: MBROLA

Στην ενότητα αυτή, θα περιγράψουμε την προέλευση, τις γενικές αρχές και την φιλοσοφία του αλγόριθμου Multi Band Resynthesis OverLap & Add (M.B.R.O.L.A), ο οποίος χρησιμοποιήθηκε στον “Αίσωπο” (E.S.o.P.o.S[®] - TTS). Αρχικώς, θα εξηγηθεί, η απλότητα και η αποτελεσματικότητα του αλγορίθμου Time Domain - Pitch Synchronous OverLap Add στον οποίο βασίζεται ο αλγόριθμος του “Αίσωπου”. Στην συνέχεια, θα παρουσιαστούν τα μειονεκτήματα της τεχνικής TD-PSOLA, και θα εξηγηθεί πώς ο τρόπος αντιμετώπισής τους με την βοήθεια αρμονικών/στοχαστικών τεχνικών, δανεισμένων από την θεωρία των Multi Band Excitation Vocoders, οδηγούν στον αλγόριθμο M.B.R.O.L.A. Ο αλγόριθμος αυτός, επινοήθηκε στο Πανεπιστήμιο της Mons (Βέλγιο), από τον καθηγητή Thierry Dutoit.

2.1 Γενική αρχή αλγόριθμου MBROLA

Ο αλγόριθμος αυτός, επιτρέπει την χρησιμοποίηση προηχογραφημένων ψηφίδων φωνής, ενώ παράλληλα επιτρέπει την μεταβολή της προσωδίας των τμημάτων (segments) αυτών. Συγκεκριμένα, είναι μια τεχνική σύνθεσης με αλυσιδωτή-διαδοχική σύνδεση ηχογραφημένων τμημάτων φωνής (**concatenation synthesis**), στα οποία όμως είναι δυνατή η αλλαγή δύο από τις τρεις προσωδικές παραμέτρους: η παράμετρος της διάρκειας (**duration**) και η παράμετρος της βασικής συχνότητας F_0 (ακουστικώς αντιλαμβανόμενη, η παράμετρος αυτή ονομάζεται **pitch**) μπορούν να μεταβληθούν ανάλογα με τις απαιτήσεις, ενώ η παράμετρος της έντασης (**intensity**) παραμένει αμετάβλητη.

2.2 Προέλευση-Φιλοσοφία

Ο αλγόριθμος M.B.R.O.L.A. ή αλλιώς Multi-Band Resynthesis Pitch Synchronous OverLap Add αποτελεί βελτίωση της Time Domain-Pitch Synchronous OverLap Add (T.D. - P.S.O.L.A.), καθώς δανείζεται τεχνικές ανασύνθεσης από το γνωστό Multi-Band Excited μοντέλο (Griffin, 1987), όπως αυτό χρησιμοποιείται σε τεχνικές κωδικοποίησης δορυφορικών σημάτων.

Η βασική ιδέα της TD-PSOLA [Charpentier & Stella (1986), Charpentier & Moulines (1989)] είναι ότι μπορούμε να παράγουμε έμφωνους (voiced) φθόγγους τροποποιημένης

βασικής συχνότητας F_0 (από εδώ και πέρα θα χρησιμοποιούμε τον όρο *pitch*), χρησιμοποιώντας τμήματα ηχογραφημένης φωνής από τα οποία έχουμε εξάγει παράθυρα βασικών περιόδων (που αντιπροσωπεύουν την γλωττιδική δραστηριότητα, το ανοιγοκλείσιμο δηλαδή των φωνητικών χορδών). Η μεταβολή του *pitch*, γίνεται αθροίζοντας τα παράθυρα βασικών περιόδων που έχουν εξαχθεί, με σχετική μεταξύ τους μετατόπιση στο χρόνο (*time sift*), έτσι ώστε η βασική περίοδος T_0 να αλλάξει στην επιθυμητή T . Η μεταβολή της διάρκειας, γίνεται με εισαγωγή ή εξαγωγή παραθύρων βασικής περιόδου.

Οι μαθηματικές σχέσεις που περιγράφουν τα παραπάνω, έχουν ως εξής:

$$s_i(n) = x(n) \cdot w(n - iT_0) \quad (1)$$

$$s(n) = \dots + s_0(n) + s_1(n - (T - T_0)) + s_2(n - 2(T - T_0)) + \dots = \sum_i s_i(n - i(T - T_0)) \quad (2)$$

Το είδος και το μήκος του παραθύρου $w(n)$, εκλέγονται προσεχτικά. Μεγάλο μήκος παραθύρου (ως προς T_0), οδηγεί σε περιοδικότητα ενδεχομένως αντιληπτή από το ανθρώπινο αυτί, με συνέπεια ένα είδος “αντήχησης” στον παραγόμενο ήχο. Από την άλλη μεριά, μικρό μήκος παραθύρου (ως προς T_0), οδηγεί **σε μικρή αναλυτικότητα συχνότητας**, με συνέπεια την φασματική αλλοίωση των εξαγόμενων παραθύρων φυσικής φωνής.

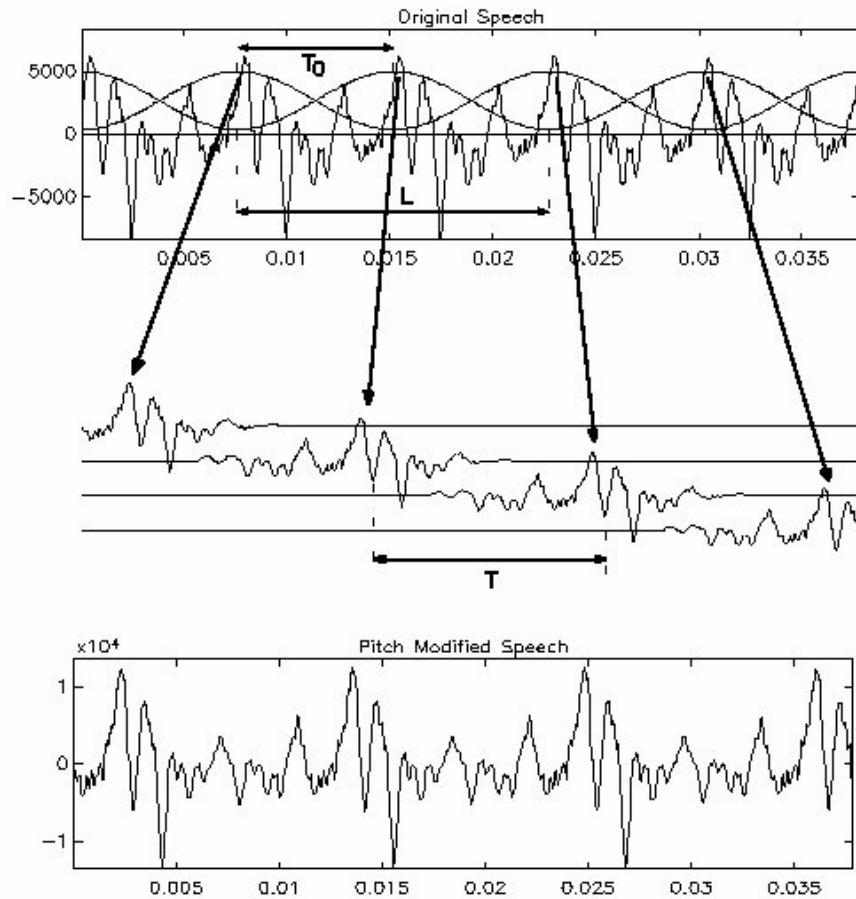
Επομένως, το μήκος των παραθύρων L αποτελεί ένα *trade off* μεταξύ των παραπάνω ιδιοτήτων και καθορίζεται από τον παράγοντα $F_R = L / T_0$ (**recovery factor**) ο οποίος εκλέγεται, συνήθως, ίσος με το δύο ($F_R = 2$). Το παράθυρο, είναι συνήθως ένα παράθυρο Hamming, καθώς αυτό το είδος των παραθύρων παρουσιάζει ικανοποιητική απόσβεση των πλευρικών λοβών του σχέση με τον κεντρικό.

Στο σχήμα που ακολουθεί, φαίνεται καθαρά η “λογική” του αλγορίθμου TD-PSOLA.

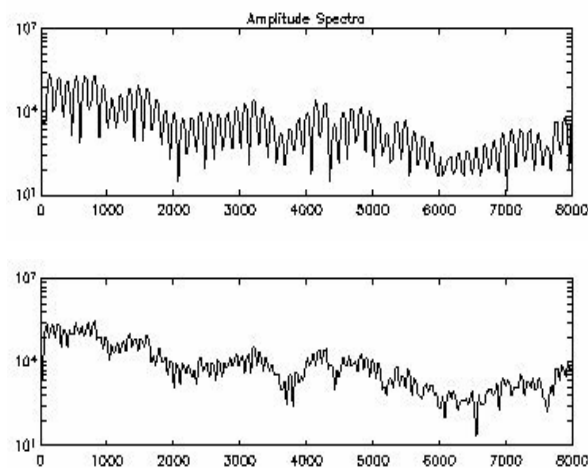
Όταν $T = T_0$, το ανακατασκευασμένο σήμα, είναι περίπου ανάλογο του αρχικού. Η αναλογία εξαρτάται από το είδος του παραθύρου που χρησιμοποιήθηκε:

$$\begin{aligned} (1) & \\ (2) & \Rightarrow s(n) = \sum_i x(n)w(n - iT) = x(n)w^*(n) \cong a \cdot x(n) \end{aligned}$$

Το παράθυρο $w(n)$, οφείλει να έχει διάρκεια αρκετά μεγάλη, έτσι ώστε το φάσμα του $w^*(n)$ να έχει μικρό εύρος (αντιστοιχώντας σε μια δέλτα του Dirac) με συνέπεια, η συνέλιξή του στο πεδίο της συχνότητας με το φάσμα του $x(n)$, να δώσει το φάσμα του $x(n)$.



[$F_R = L/T_0 = 2$, αφού εξαχθούν 2 βασικοί περίοδοι ανά παράθυρο, κάθε παράθυρο μετατοπίζεται κατά $T - T_0$, έτσι ώστε να επιτευχθεί κατά την άθροιση, περίοδος T . (Dutoit & Leich, 1993)]



[στο επάνω σχήμα φαίνεται το φάσμα του αρχικού σήματος, περιόδου T_0 , ενώ στο κάτω σχήμα φαίνεται το φάσμα του ανακατασκευασμένου σήματος, περιόδου $T \neq T_0$ (Dutoit & Leich, 1993)]

Η ποιότητα των τμημάτων που παράγονται με την τεχνική TD-PSOLA είναι εξαιρετική, όταν η τελευταία εφαρμόζεται σε ημιστάσιμα (quasi-stationary) τμήματα του λόγου (καθώς, όπως είναι γνωστό, η φωνή δεν είναι ποτέ στάσιμο σήμα).

Ωστόσο, τα μεγαλύτερα προβλήματα της TD-PSOLA, όπως και των περισσότερων concatenation τεχνικών, εμφανίζονται κατά την σύνδεση τμημάτων φωνής που έχουν εξαχθεί από διαφορετικά λεκτικά-ακουστικά περιβάλλοντα (π.χ. διαφορετικές λέξεις), εξαιτίας τριών, σχεδόν αναπόφευκτων, γεγονότων:

1. *Ασυνέχειες φάσης, κατά την εξαγωγή των παραθύρων (Phase mismatch)*
2. *Διαφορετικό pitch, των ηχογραφημένων έμφωνων φθόγγων (Pitch mismatch)*
3. *Ασυνέχειες φασματικού φακέλου (Spectral envelope mismatch)*

Ας αναλύσουμε, κάθε περίπτωση ξεχωριστά:

1. Phase mismatch

Ακόμη και εάν οι προς σύνδεση κυματομορφές είναι παρόμοιες, οι λειτουργίες της επικάλυψης και της πρόσθεσης (OverLap & Add, OLA) μπορούν να οδηγήσουν σε απρόσμενες ασυνέχειες, αν τα διαδοχικά παράθυρα δεν τοποθετηθούν στο ίδιο σημείο της βασικής περιόδου.

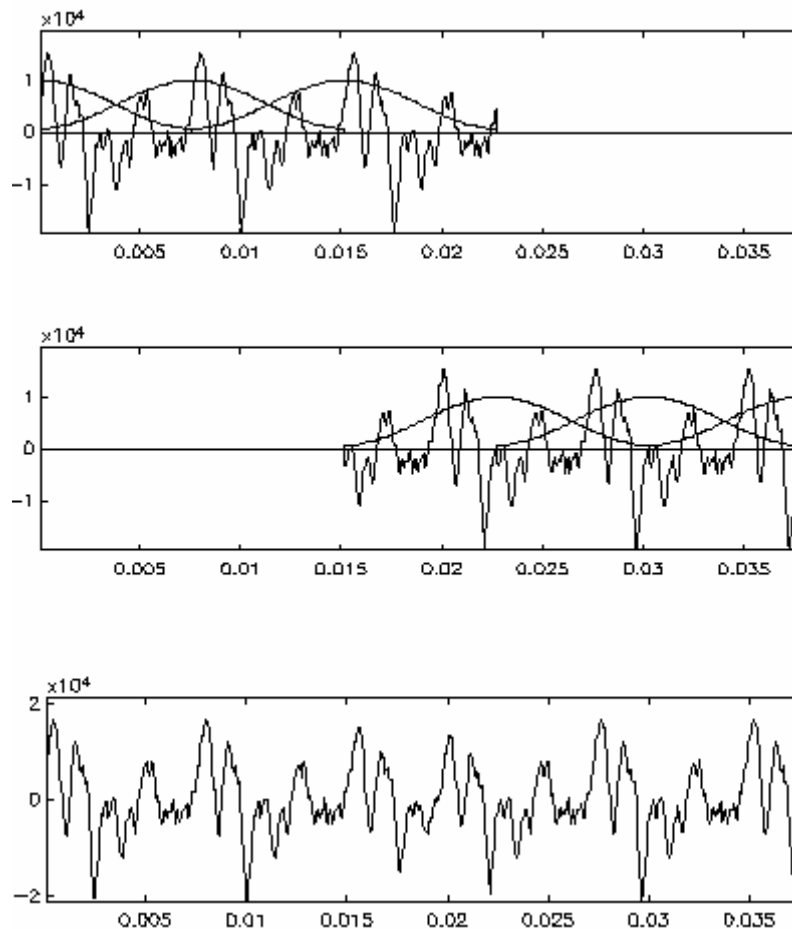
Η τοποθέτηση του μέσου κάθε παραθύρου στο σημείο (glottis closure) που κλείνει η γλωττίδα¹, με την χρησιμοποίηση ενός οπτικού περιβάλλοντος επεξεργασίας φωνής, αποτελεί μια λύση με ακρίβεια, ωστόσο είναι ιδιαιτέρως επίπονη και χρονοβόρα.

Επιπλέον, οποιαδήποτε αυτόματη τεχνική “σύμφωνης” (coherent) εξαγωγής των παραθύρων που έχει χρησιμοποιηθεί, οδηγεί σε σημαντικά σφάλματα φάσης, όταν δεν συνοδεύεται από παράλληλη επίβλεψη. Αυτό συμβαίνει εξαιτίας της ποικιλίας των ανθρώπινων φθόγγων. Οποιαδήποτε τεχνική αυτόματης, σύμφωνης εξαγωγής των βασικών περιόδων οφείλει να λαμβάνει υπόψιν για παράδειγμα, το μακρύ ανέβασμα και κατέβασμα των ρινικών (nasals) και κροτικών (plosives) φθόγγων που αντιστοιχεί στο low-pass φιλτράρισμα από τον φωνητικό σωλήνα (vocal tract) των γλωττιδικών παλμών. Ταυτόχρονα, η τεχνική θα πρέπει να λαμβάνει υπόψιν τον υψίσυχνο χαρακτήρα των προστριβόμενων φθόγγων (fricatives) που οφείλεται στις μεγάλες συχνότητες διαμόρφωσης (formants) των φθόγγων αυτών.

Συνεπώς, οποιαδήποτε “γρήγορη” εξαγωγή των παραθύρων, θα οδηγήσει σε σφάλματα, τα οποία με την σειρά τους θα οδηγήσουν σε σφάλματα φάσης. Ίσως, στην προσεκτική και σύμφωνη εξαγωγή των παραθύρων να βοηθούσε ένας ηλεκτρο-λαρυγγογράφος, ο

¹ το ανοίγμα που σχηματίζουν οι φωνητικές χορδές

οποίος καταγράφει αυτόματα την γλωττιδική δραστηριότητα, χωρίς επομένως να χρειάζεται έμμεση καταγραφή της από το σήμα της φωνής (Conkie, 1996).



[οι κυματομορφές έχουν ίδιο pitch, ωστόσο για $F_R=2$, τα παράθυρα έχουν εξαχθεί “ασύμφωνα”]

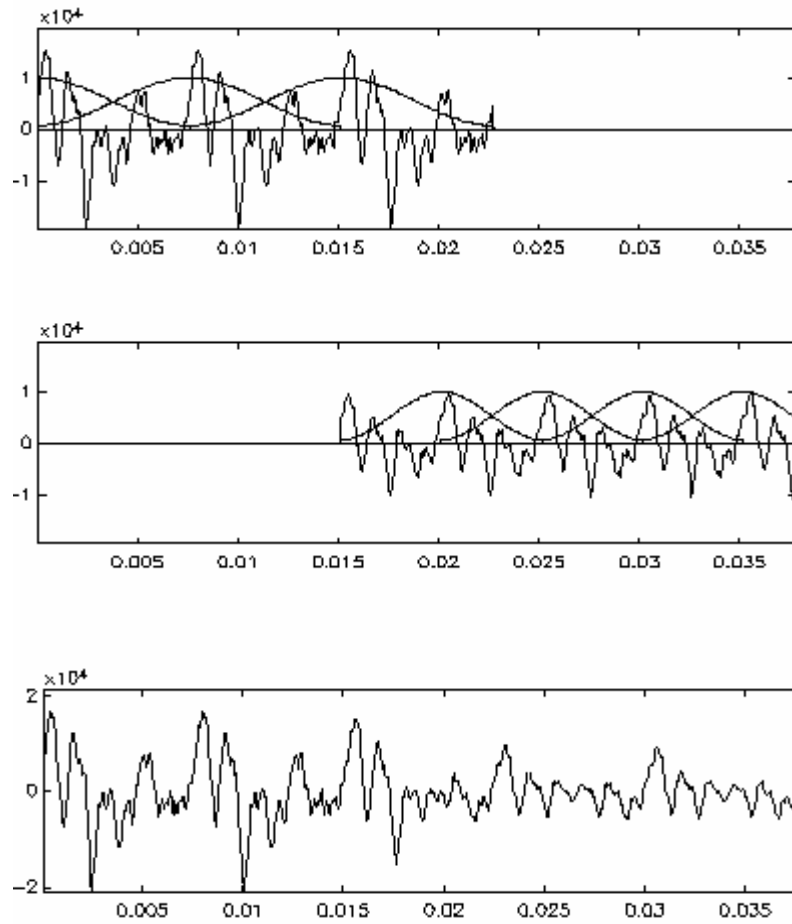
2. Pitch mismatch

Η δυνατότητα αλλαγής της βασικής συχνότητας, μέσω επικάλυψης και πρόσθεσης βασικών περιόδων, προϋποθέτει την σταθερότητα του μήκους των περιόδων αυτών. Ωστόσο, κατά την ηχογράφηση των ψηφίδων φωνής που θα χρησιμοποιηθούν, είναι αδύνατον ο εκφωνητής (εκπαιδευμένος ή όχι) να διατηρήσει 100% σταθερό pitch. Συνεπώς στην βάση των προηχογραφημένων ψηφίδων φωνής ενυπάρχει κάποιο ποσοστό φωνής με pitch που αποκλίνει από την μέση τιμή.

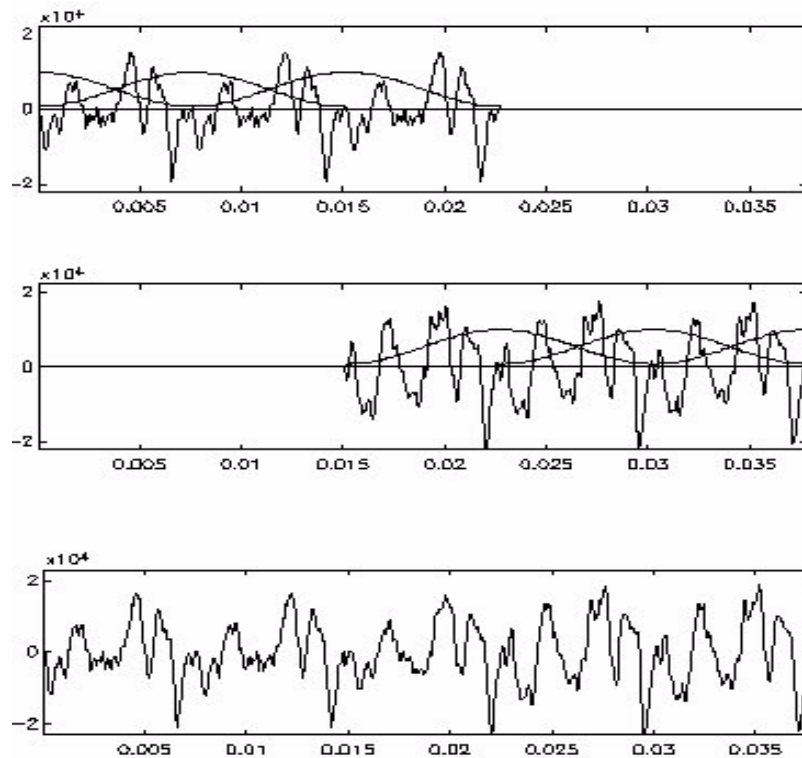
3. Spectral envelope mismatch

Ο διαφορετικός φασματικός φάκελος των προς σύνδεση τμημάτων, οφείλεται κυρίως, στα φαινόμενα συνάρθρωσης, που υπάρχουν εγγενώς σε κάθε γλώσσα, οδηγώντας σε ανεπιθύμητα “clicks” κατά την σύνδεση. Η TD-PSOLA, όπως και οι περισσότερες

concatenation τεχνικές, δεν μπορούν να εξομαλύνουν τους φασματικούς φακέλους στα σημεία ένωσης (concatenation points).



[τα OLA παράθυρα έχουν εξαχθεί “σύμφωνα”, από διαφορετικού, όμως, pitch τμήματα]



[το επάνω τμήμα είναι το δίφωνο 'ma' ενώ το μεσαίο τμήμα, είναι το δίφωνο 'am'. Τα παράθυρα έχουν εξαχθεί σύμφωνα, ενώ το pitch είναι σταθερό. Η φασματική ασυνέχεια, επικεντρώνεται σε μια περίοδο.

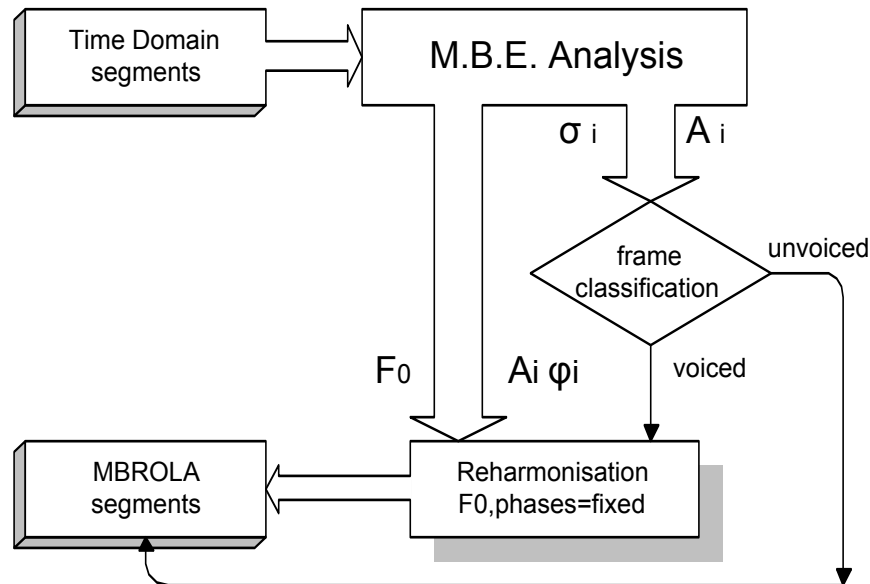
(Dutoit & Leich, 1993)]

2.3 Δυνατότητες MBROLA τεχνικής

Είναι επομένως κατανοητές οι αδυναμίες της τεχνικής TD-PSOLA. Ενώ η φιλοσοφία της είναι ιδιαίτερος απλή και έξυπνη, αδυνατεί να αποφύγει τις “ενοχλητικές” ασυνέχειες στα σημεία ένωσης, που οφείλονται σε μια, τουλάχιστον, από τις παραπάνω αιτίες.

Ο αλγόριθμος Multi Band Resynthesis OverLap & Add έρχεται να δώσει λύσεις στα παραπάνω προβλήματα, καθώς ανακατασκευάζει το σύνολο των ψηφίων φωνής που θα συνδεθούν αλυσιδωτά, με βάση την τεχνική TD-PSOLA. Η ανασύνθεση της βάσης των τμημάτων φωνής, θα οδηγήσει σε ένα σύνολο έμφωνων φθόγγων:

- α. με σταθερό pitch
- β. με σύμφωνη εξαγωγή των παραθύρων
 βασικών περιόδων
- γ. με δυνατότητα φασματικής εξομάλυνσης
 στα σημεία συνένωσης



Ο αλγόριθμος MBROLA ή Multi Band Resynthesis - Pitch Synchronous OverLap & Add, επινοήθηκε από τους Dutoit και Leich το 1993, και ανασυνθέτει τις ψηφίδες φωνής που θα χρησιμοποιηθούν, υλοποιώντας ένα παραμετρικό, αρμονικό/στοχαστικό μοντέλο, βασισμένο στο αντίστοιχο μοντέλο που χρησιμοποιείται στους Multi Band Excitation Vocoders (Griffin, 1987).

Το πλαίσιο ανάλυσης, καθορίζεται από την μέση βασική περίοδο (το αντίστροφο του pitch) των ψηφίδων φωνής, η οποία υπολογίζεται με μεγάλη ακρίβεια με βάση την M.B.E. ανάλυση. Το αρμονικό/στοχαστικό μοντέλο που ακολουθείται, υποθέτει πως κάθε φθόγγος μπορεί να εκφραστεί ως ένα άθροισμα μιας βασικής συχνότητας και των αρμονικών της, με narrow band θόρυβο, σύμφωνα με την σχέση :

$$\tilde{s}(t) = \sum_i a_i(t) \cos(\phi_i(t)) + \sum_j \sigma_j(t) \eta_0(t) \cos(\phi_j(t)), \text{ όπου } \eta_0 \text{ είναι low passed θόρυβος}$$

μοναδιαίας μεταβλητότητας.

Στο στάδιο ανάλυσης, υπολογίζονται τα a_i (A_i), ϕ_i, σ_i και αφού διαχωριστούν τα τμήματα, σε έμφωνα και άφωνα, γίνεται ανασύνθεση μόνο για τα έμφωνα τμήματα.

Επομένως, τα έμφωνα τμήματα ανασυνθέτονται με pitch, ίσο με το μέσο pitch που βρέθηκε στο στάδιο της ανάλυσης, λαμβάνοντας υπόψιν τον φασματικό φάκελο του κάθε τμήματος, όπως υπολογίστηκε στο στάδιο της ανάλυσης. Για κάθε βασική περίοδο, οι φάσεις διατηρούνται σταθερές μόνο για τις υψηλότερες αρμονικές, ενώ για τις χαμηλότερες αρμονικές, ορίζονται σε τυχαίες αλλά συγκεκριμένες τιμές. Ο ορισμός των φάσεων όλων των αρμονικών σε τυχαίες αλλά συγκεκριμένες τιμές, οδηγεί σε

“μεταλλικούς” ήχους, γι’ αυτό στην MBROLA τεχνική, οι φάσεις γίνονται reset μέχρι τα 2 kHz, ενώ για τις υψηλότερες αρμονικές διατηρούνται οι αρχικές τιμές που υπολογίστηκαν στο στάδιο της ανάλυσης.

Προκύπτουν επομένως, σταθερού pitch κυματομορφές, στις οποίες πλέον είναι πολύ εύκολη η σύγχρονη εξαγωγή των παραθύρων (με $F_R=2$).

Η σταθερή βασική συχνότητα (pitch) και η διαδικασία reset των φάσεων που οδηγεί στις ίδιες φάσεις των αρμονικών για κάθε βασική περίοδο των επεξεργασμένων τμημάτων φωνής, καθιστά κατανοητή την *ισοδυναμία της παρεμβολής-εξομάλυνσης του φασματικού φακέλου των προς ένωση τμημάτων (spectral envelope interpolation) με την απευθείας εξομάλυνσή τους, στο πεδίο του χρόνου (temporal interpolation)* (Dutoit & Leich et. al.)

Η διαδικασία ανασύνθεσης, οδηγεί επομένως σε ψηφίδες φωνής σταθερού pitch που με την διαδικασία reset των φάσεων, επιτρέπει την φασματική εξομάλυνση διαφορετικών ψηφίδων φωνής, κατά την αλυσιδωτή σύνδεσή τους, με μια γραμμική διαδικασία στο πεδίο του χρόνου.

Οι μαθηματικές σχέσεις που χρησιμοποιούνται για την εξομάλυνση αυτή περιγράφονται παρακάτω.

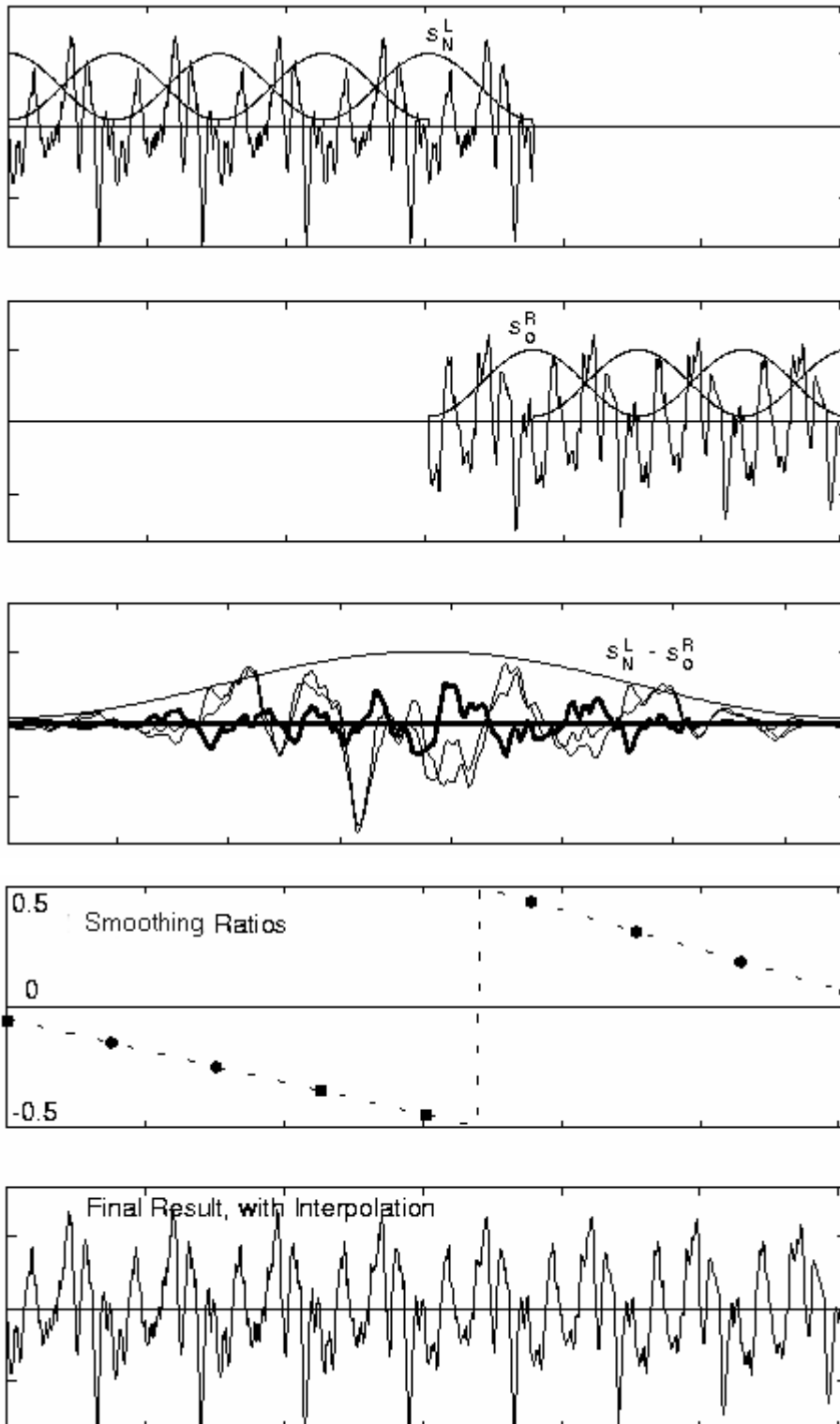
Για τα άφωνα (unvoiced) τμήματα της φωνής, η αλλαγή της διάρκειας γίνεται με προσθαφαίρεση προκαθορισμένων τμημάτων τους, ενώ για να αποφευχθούν ανεπιθύμητες “αντηχήσεις” από την πρόσθεση και επομένως επανάληψη συγκεκριμένης κυματομορφής, τα τμήματα που προστίθενται, προστίθενται με αντιστροφή του άξονα του χρόνου!

$$s_{N-i}^L = s_{N-i}^L + (s_0^R - s_N^L) \frac{1}{2} \left(\frac{M_L - i - 0.5}{M_L} \right)$$

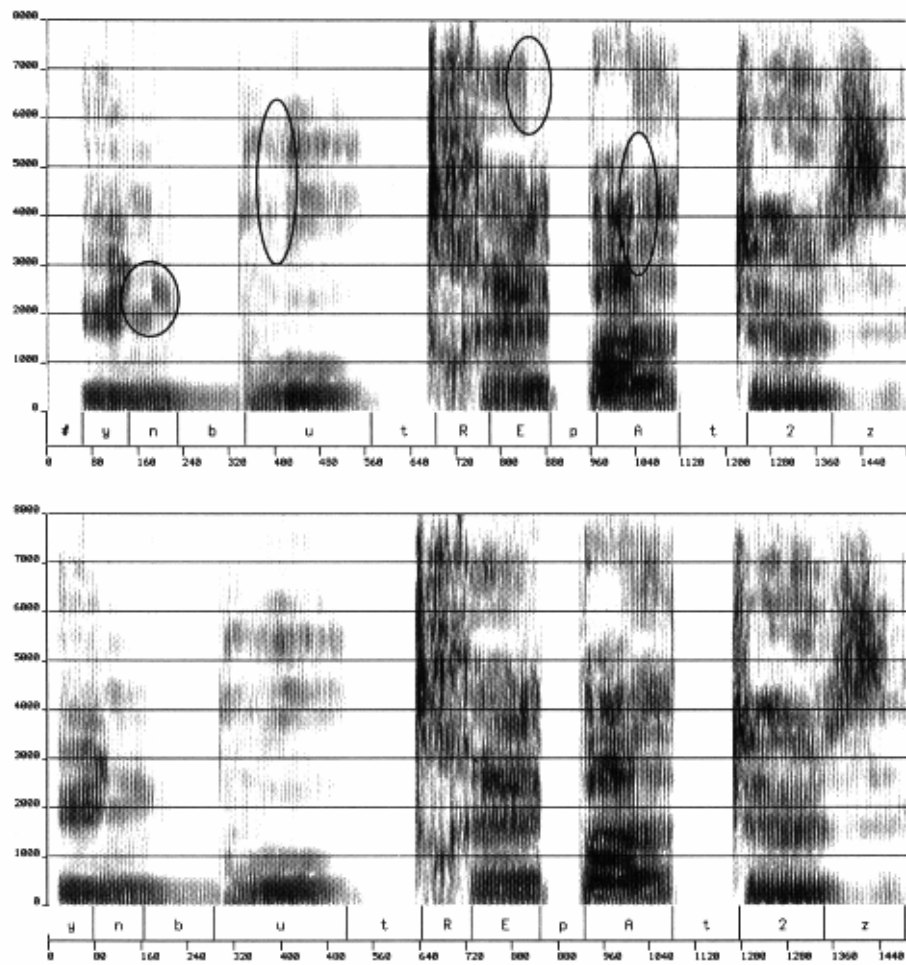
$$s_j^R = s_j^R + (s_N^L - s_0^R) \frac{1}{2} \left(\frac{M_R - i - 0.5}{M_R} \right)$$

Αν s_N^L , s_0^R συμβολίζουν αντιστοίχως, το τελευταίο από αριστερά OLA (OverLap Add) παράθυρο και το πρώτο από δεξιά παράθυρο, τότε πρέπει να τροποποιήσουμε έναν συγκεκριμένο αριθμό M_L και M_R από OLA παράθυρα πριν από το s_N^L και μετά από το s_0^R σύμφωνα με τους παραπάνω τύπους. Ουσιαστικά, η διαφορά των προς σύνδεση τμημάτων s_N^L & s_0^R , μεταφέρεται σε ένα αριθμό M_L και M_R τμημάτων, αριστερά και δεξιά αντιστοίχως, του σημείου σύνδεσης με μεγαλύτερο βάρος (smoothing ratio) στα τμήματα πιο κοντά στο σημείο ένωσης (concatenation point).

Η έξυπνη αυτή φασματική εξομάλυνση, που ωστόσο γίνεται στο πεδίο του χρόνου, φαίνεται στα παρακάτω σχήματα:



[Εξομάλυνση με $F_R=2$, $M_L=5$ και $M_R=4$ (Dutoit & Leich)]



[φαίνεται καθαρά, η ικανότητα φασματικής εξομάλυνσης του αλγορίθμου MBROLA]

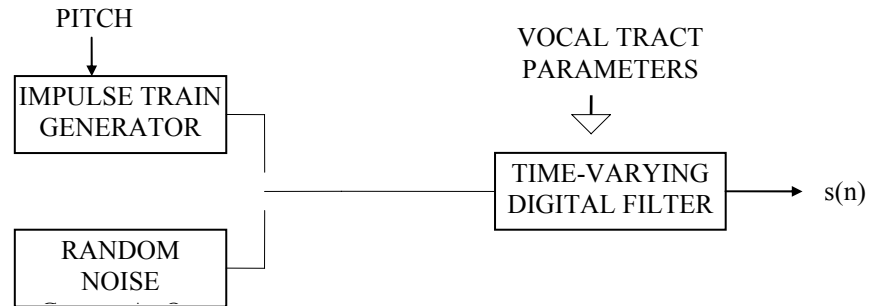
μέρος δεύτερο: Κωδικοποίηση Γραμμικής Πρόβλεψης L.P.C.

Στην ενότητα αυτή, θα περιγράψουμε την βασική αρχή που διέπει την θεωρία Γραμμικής Πρόβλεψης (Linear Prediction) και θα την εφαρμόσουμε, με χρήση ορισμένων μεθόδων, θέλοντας να καταδείξουμε την χρησιμότητα της ως βασικό εργαλείο ανάλυσης του σήματος της φωνής.

Αφού κατανοήσουμε την φιλοσοφία της, θα προχωρήσουμε στον συσχετισμό της θεωρίας L.P.C με υπάρχουσες και ευρέως εφαρμοσμένες τεχνικές Σύνθεσης Φωνής.

2.4 LPC : Γενική αρχή - μέθοδος Ανάλυσης της Φωνής

Το ευρέως αποδεκτό μοντέλο σύνθεσης φωνής, προσομοιώνει τον φωνητικό σωλήνα (vocal tract)¹ με ένα γραμμικό, χρονομεταβλητό ψηφιακό φίλτρο που διεγείρεται από μια σειρά ώσεων ή λευκό θόρυβο, για την παραγωγή έμφωνων ή άφωνων φθόγγων αντίστοιχα.



Η θεωρία Γραμμικής Πρόβλεψης υποθέτει πως κάθε δείγμα του σήματος της φωνής, μπορεί να προβλεφθεί από τα προηγούμενα δείγματα με βάση τον τύπο (1):

$$\tilde{s}(n) = \sum_{k=1}^p a_k s(n-k) \quad (1), \quad e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^p a_k s(n-k) \quad (2)$$

όπου οι συντελεστές a_k λέγονται συντελεστές πρόβλεψης. Το σφάλμα πρόβλεψης υπολογίζεται από τον τύπο (2). Επομένως, το φίλτρο με χαρακτηριστική συνάρτηση $A(z)$ και είσοδο το ίδιο το σήμα $s(n)$, παράγει το σήμα σφάλματος $e(n)$, γι'αυτό το σήμα σφάλματος λέγεται και εναπομένον σήμα αντίστροφου φιλτραρίσματος (residual part).

$$A(z) = 1 - \sum_{k=1}^p a_k z^{-k}$$

Η θεωρία Γραμμικής Πρόβλεψης, θεωρεί πως το ψηφιακό φίλτρο που αντιπροσωπεύει τον φωνητικό σωλήνα, είναι της μορφής $1/A(z)$, είναι δηλαδή ένα all pole φίλτρο, τάξης p .

Συνεπώς, το μοντέλο παραγωγής φωνής, σύμφωνα με την LPC θεωρία, περιγράφεται από τους παρακάτω τύπους:

$$H(z) = \frac{S(z)}{D(z)} = \frac{1}{1 - \sum_{k=1}^p a_k z^{-k}} = \frac{1}{A(z)} \quad (\text{I.I.R. ευθεία μορφή τύπου I}) \quad (3)$$

$$e(n) = s(n) - \sum_{k=1}^p a_k s(n-k) \quad (4)$$

$$(3), (4) \Rightarrow e(n) = d(n) \quad (5),$$

όπου $d(n)$ είναι το σήμα διέγερσης στο φίλτρο του φωνητικού σωλήνα, και αντιπροσωπεύει το ρεύμα αέρα που εκπνέεται από τους πνεύμονες και εξέρχόμενο από τον φωνητικό σωλήνα, διαμορφώνεται κατάλληλα, με συνέπεια την παραγωγή του επιθυμητού φθόγγου.

Η θεωρία Γραμμικής Πρόβλεψης επομένως, υποθέτει πως το σήμα της φωνής, είναι προϊόν συνέλιξης δύο ανεξάρτητων τμημάτων: του φίλτρου φωνητικού σωλήνα με το σήμα διέγερσης $d(n)$, το οποίο ταυτίζεται με το σήμα σφάλματος πρόβλεψης $e(n)$! Η υπόθεση της συνέλιξης, απαιτεί **γραμμικότητα** και **χρονοαμεταβλητότητα** για το φίλτρο του φωνητικού σωλήνα (vocal tract). Η απαίτηση αυτή ικανοποιείται για μικρά χρονικά διαστήματα, όταν οι συντελεστές πρόβλεψης a_k εγκλωβίζουν χρήσιμες πληροφορίες για τις συχνότητες διαμόρφωσης (formants), steady-state φθόγγων.

Το σήμα διέγερσης (σήμα σφάλματος), από την άλλη μεριά, **δεν** περιέχει πληροφορία σχετικά με τις formants, αλλά περιγράφει τους γλωττιδικούς παλμούς² (glottal pulses) σε συνδυασμό με τον απαραίτητο συνοδευτικό “θόρυβο” (aspiration noise) στην περίπτωση των έμφωνων φθόγγων, ή τον υψίσυχνο “θόρυβο” στην περίπτωση των άφωνων φθόγγων.

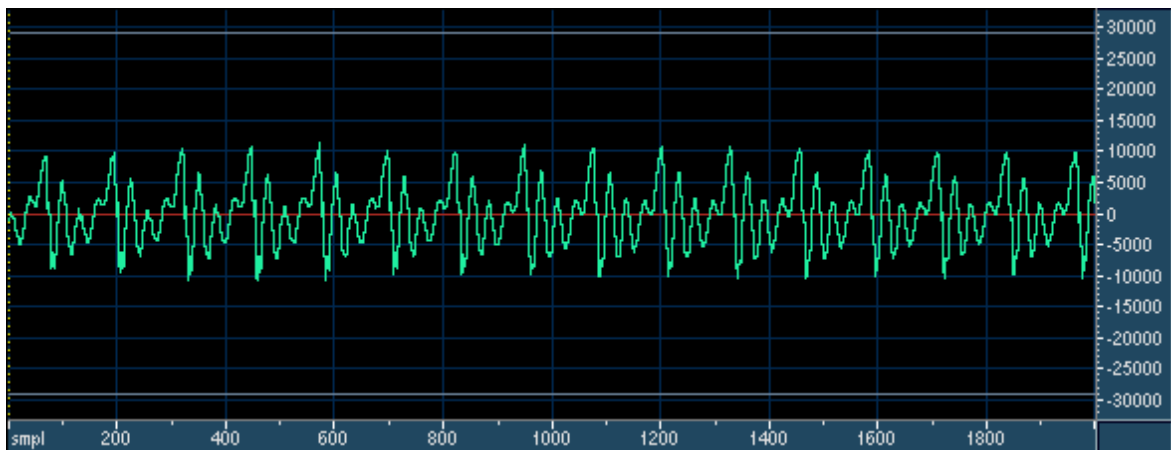
Η τάξη p του all-pole φίλτρου, οφείλει να είναι όσο το δυνατόν μεγαλύτερη, ωστόσο για λόγους υπολογιστικού κόστους, η τιμή της περιορίζεται. Τόσο οι Rabiner & Schafer (Rabiner & Schafer, 1978), όσο και οι Markel & Gray (Markel & Gray, 1976), αναφέρουν πως σε κάθε kHz συχνότητας δειγματοληψίας αντιστοιχεί ένα ζευγάρι μιγαδικών πόλων, συνεπώς για συχνότητα δειγματοληψίας F_s (σε kHz) αντιστοιχούν F_s πόλοι. Σ’ αυτούς θα πρέπει να προσθέσουμε και έναν μικρό αριθμό πόλων, για την ικανή αντιπροσώπευση του φάσματος της διέγερσης (source excitation spectrum) και του φαινομένου διάδοσης (radiation load). Για τα 16 kHz συχνότητας δειγματοληψίας της

¹ ο φωνητικός σωλήνας, θεωρούμε πως ξεκινά από την γλωττίδα, περιλαμβάνει τον φάρυγγα και καταλήγει στην στοματική κοιλότητα που συνδέεται με την ρινική κοιλότητα

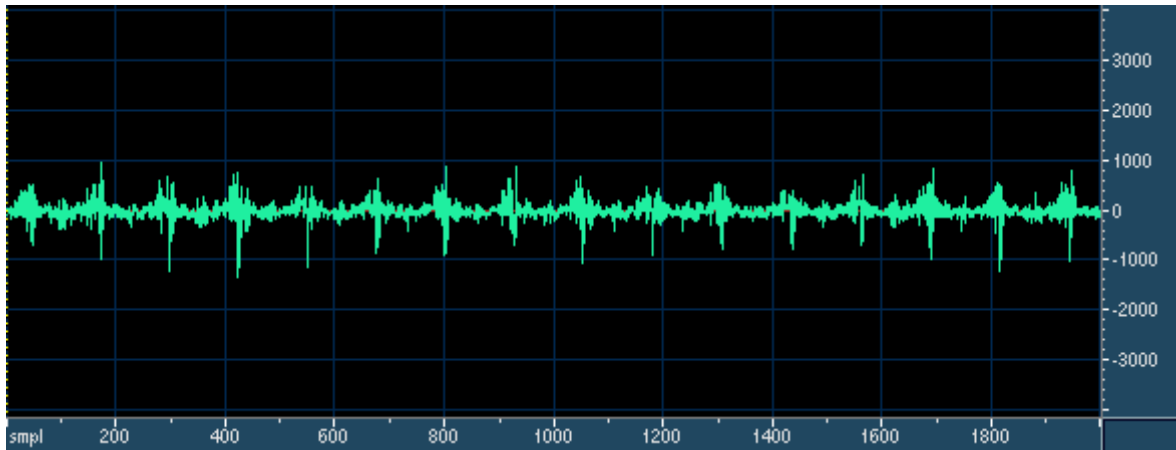
² οι παλμοί πίεσης του αέρα που προκύπτουν από το ανοιγοκλείσιμο των φωνητικών χορδών

εργασίας αυτής, χρησιμοποιήθηκε $p=16+8=24$. Για μεγαλύτερο p , τα αποτελέσματα δεν μεταβλήθηκαν.

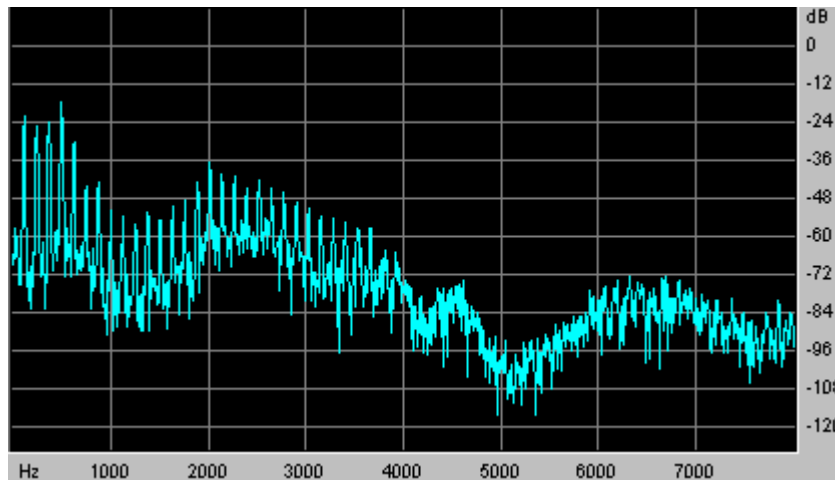
Ακριβώς, επειδή το παραπάνω μοντέλο εφαρμόζεται ανά τμήματα φωνής, λόγω της χρονικής μεταβλητότητας του σήματος, υπάρχει η ανάγκη χρησιμοποίησης παραθύρου. Το μήκος του N θα πρέπει να είναι επαρκώς μεγάλο, έτσι ώστε η αναλυτικότητα συχνότητας να είναι μεγάλη, χωρίς επομένως το παράθυρο να αλλοιώνει τα φασματικά χαρακτηριστικά του σήματος της φωνής, με την οποία πολλαπλασιάζεται.



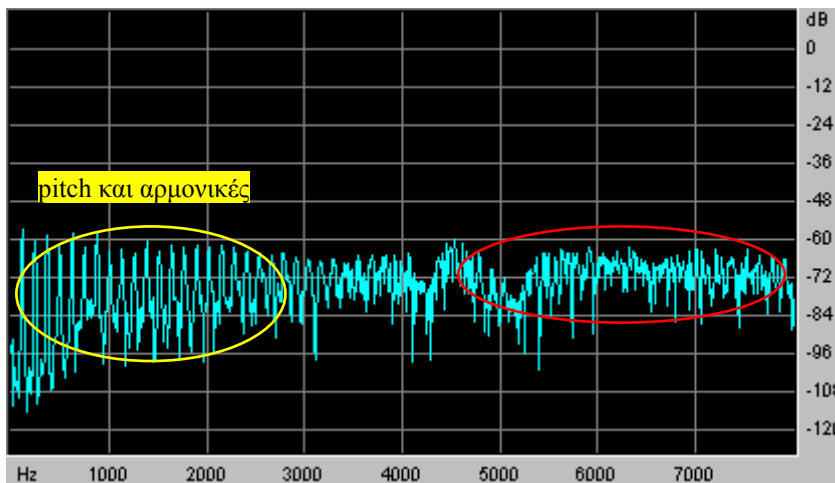
[2000 δείγματα με $F_s=16\text{kHz}$ steady state /e/]



[σήμα σφάλματος $e(n)$ (residual) από αντίστροφο φιλτράρισμα του παραπάνω σήματος /e/ (μήκος παραθύρου $N=2000$, τετραγωνικό παράθυρο, $p=24$)]



[το αρχικό /e/ στο πεδίο Fourier. Φαίνεται καθαρά η βασική συχνότητα (pitch) και οι αρμονικές της, ενώ η περιβάλλουσα, απεικονίζει την διάταξη των formants]

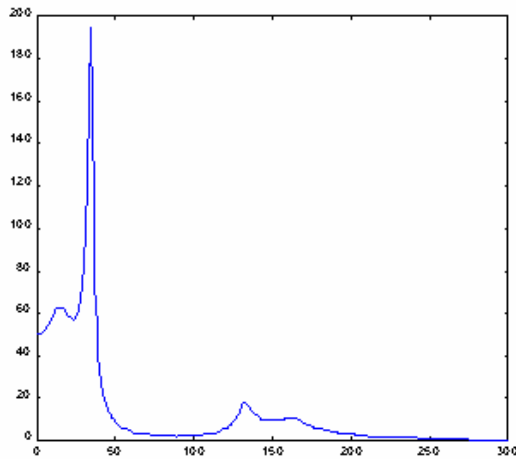


[το σήμα σφάλματος $e(n)$ στο πεδίο Fourier. Φαίνεται καθαρά το pitch και οι αρμονικές του, ενώ προφανής είναι η απουσία των formants, η πληροφορία των οποίων έχει “κρατηθεί” στο φίλτρο $1/A(z)$ του φωνητικού σωλήνα. Επίσης, φαίνεται στο υψίσυχο τμήμα του φάσματος, ο συνοδευτικός “θόρυβος” (π.χ. aspiration noise)]

Το φίλτρο $H(z) = \frac{S(z)}{D(z)} = \frac{1}{1 - \sum_{k=1}^p a_k z^{-k}} = \frac{1}{A(z)}$, εγκλωβίζει μέσω των συντελεστών a_k

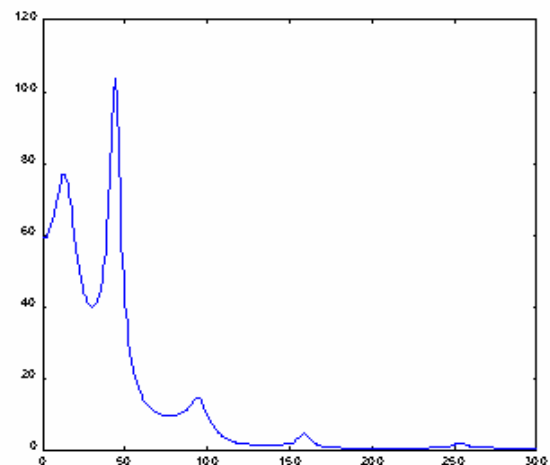
χρήσιμες πληροφορίες για τις συχνότητες διαμόρφωσης (formants) του κάθε φθόγγου. Επομένως, η σχεδίαση της απόκρισης συχνότητας³ του φίλτρου και ο επακόλουθος υπολογισμός των formants, είναι πολύ εύκολος εφόσον βρεθούν τα a_k . Για το πώς υπολογίζονται οι συντελεστές πρόβλεψης a_k θα μιλήσουμε παρακάτω.

³ Η απόκριση συχνότητας, βρίσκεται για $z=e^{j\omega}$, με την προϋπόθεση ότι το φίλτρο είναι ευσταθές. Τα a_k στην εργασία αυτή υπολογίζονται με την μέθοδο της αυτοσυσχέτισης που οδηγεί σε ευσταθές φίλτρο.



[απόκριση συχνότητας vocal tract για το /e/]

formants/phones	/e/
f1 (Hz)	234
f2 (Hz)	531
f3 (Hz)	2078



[απόκριση συχνότητας vocal tract για το /a/]

/a/	/s/
203	2031
688	3078
1469	4734

2.5 Μέθοδοι Υπολογισμού Συντελεστών πρόβλεψης α_k

Οι συντελεστές α_k για ένα τμήμα φωνής, μπορούν να υπολογιστούν με δύο παρόμοιες μεθόδους που προϋποθέτουν υπολογισμό της αυτοσυσχέτισης (μικρού χρόνου) στο τμήμα αυτό:

1. μέθοδος μεταβλητότητας (covariance method)
2. μέθοδος αυτοσυσχέτισης (autocorrelation method)

Ουσιαστικά, οι δύο μέθοδοι διαφοροποιούνται στα όρια του τμήματος φωνής στο οποίο επεξεργάζονται. Η 1η μέθοδος, οδηγεί σε συντελεστές α_k που δίνουν καλύτερες προβλέψεις για το σήμα της φωνής, ωστόσο η μέθοδος αυτή, δεν εγγυάται την ευστάθεια του φίλτρου $H(z)$. Αντιθέτως, η 2η μέθοδος εγγυάται την ευστάθεια του φίλτρου που προσομοιώνει τον φωνητικό σωλήνα, συνεπώς αυτή ευρέως χρησιμοποιείται.

Παρακάτω θα παρουσιάσουμε τον αλγόριθμο υπολογισμού των α_k με την μέθοδο της αυτοσυσχέτισης, όπως εφαρμόστηκε στην εργασία αυτή. Για περισσότερες επεξηγήσεις, ο ενδιαφερόμενος μπορεί να αναφερθεί στην σχετική βιβλιογραφία (Markel & Gray, 1986).

Η ελαχιστοποίηση του τετραγωνικού σφάλματος $\sum_j e_n^2(j) = \sum_j [s_n(j) - \tilde{s}_n(j)]^2$, όπου

$s_n(m) = s(n+m)$ $w(m)$, είναι το τμήμα της φωνής που μας ενδιαφέρει και $w(m)$ είναι το

παράθυρο μήκους N που χρησιμοποιούμε, οδηγεί στην επίλυση ενός γραμμικού $p \times p$ συστήματος εξισώσεων:

$$\begin{bmatrix} R_n(0) & R_n(1) & R_n(2) & \dots & R_n(p-1) \\ R_n(1) & R_n(0) & R_n(1) & \dots & R_n(p-2) \\ R_n(2) & R_n(1) & R_n(0) & \dots & R_n(p-3) \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ R_n(p-1) & R_n(p-2) & R_n(p-3) & \dots & R_n(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \dots \\ \dots \\ a_p \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \\ R_n(3) \\ \dots \\ \dots \\ R_n(p) \end{bmatrix}$$

$$\text{ή πιο συνοπτικά: } \sum_{k=1}^p a_k R_n(|i-k|) = R_n(i), \quad 1 \leq i \leq p$$

$$\text{όπου } R_n \text{ η αυτοσυσχέτιση μικρού χρόνου: } R_n(k) = \sum_{m=0}^{N-1-k} s_n(m)s_n(m+k)$$

Ο παραπάνω πίνακας των συντελεστών αυτοσυσχέτισης, είναι ένας συμμετρικός Toeplitz πίνακας και μπορεί να λυθεί πολύ γρήγορα με την αναδρομική μέθοδο του Durbin :

$$E^{(0)} = R(0)$$

$$k_i = \left(R(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j) \right) / E^{(i-1)}, \quad 1 \leq i \leq p$$

$$a_i^{(i)} = k_i$$

$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)}, \quad 1 \leq j \leq i-1$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)}$$

Οι παραπάνω εξισώσεις λύνονται διαδοχικά για $i=1, 2, \dots, p$ και ο συντελεστής πρόβλεψης προκύπτει από :

$$a_j = a_j^{(p)}, \quad 1 \leq j \leq p$$

Οι συντελεστές k_i , $1 \leq i \leq p$, ονομάζονται συντελεστές ανάκλασης (reflection coefficients), καθώς η μέθοδος της αυτοσυσχέτισης αποδεικνύεται ανάλογη με την προσομοίωση του φωνητικού σωλήνα με έναν ακουστικό σωλήνα (acoustical tube) με μεταβλητή διατομή.

Αποδεικνύεται επίσης, πως εάν $|k_i| < 1$, τότε το φίλτρο $H(z)$ που προκύπτει είναι ευσταθές.

Με την μέθοδο της αυτοσυσχέτισης, οι συντελεστές ανάκλασης είναι πάντοτε απολύτως μικρότεροι της μονάδας, συνεπώς **η μέθοδος οδηγεί πάντοτε σε ευσταθή αποτελέσματα.**

2.6 LPC και Σύνθεση Φωνής

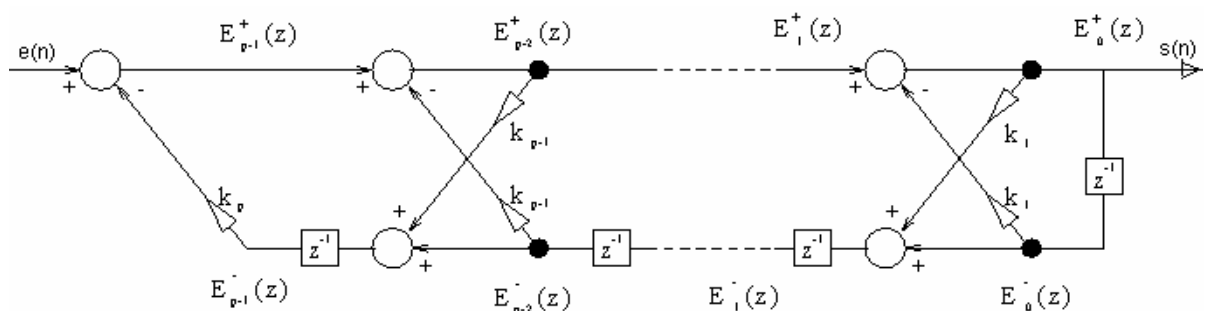
Από τα παραπάνω καταλαβαίνουμε πως το ψηφιακό φίλτρο $H(z)$ προσομοιώνει τον φασματικό φάκελο του σήματος της φωνής, ενώ το σήμα σφάλματος $e(n)$ περιέχει (στην ιδανική περίπτωση) την αρμονική δομή της φωνής ή/και λευκό θόρυβο.

Επομένως, η φωνή μπορεί να ανασυνθεθεί σε διαφορετική βασική συχνότητα ή/και διάρκεια, διεγείροντας το φίλτρο-vocal tract $H(z)$ με ένα συνθετικό (δηλαδή φτιαγμένο από εμάς) σήμα σφάλματος $e'(n)$, παρέχοντας έτσι νέο pitch ή/και νέα διάρκεια.

Η παραπάνω απλή λογική της Σύνθεσης Φωνής με LPC χρησιμοποιείται σε concatenation υλοποιήσεις, όπου προηχογραφημένα τμήματα-παράθυρα φωνής, διάρκειας περίπου 30 msec, αναλύονται στους συντελεστές πρόβλεψης, με επικάλυψη παραθύρων (time shift) 6-10 msec. Συνθετικό σήμα διέγερσης $e(n)$, καθορίζει τα επιθυμητά προσωδικά χαρακτηριστικά της παραγόμενης φωνής.

Η υλοποίηση της θεωρίας γραμμικής πρόβλεψης σε εφαρμογές Σύνθεσης, δεν γίνεται με την μορφή I.I.R. (Infinite Impulse Response) φίλτρου (ευθείας μορφής τύπου I), όπως την αναλύσαμε στην προηγούμενη παράγραφο, αλλά με την χρήση των συντελεστών ανάκλασης k_i , σε φίλτρο πλεγματικής μορφής (lattice filter), όπως φαίνεται στο σχήμα που ακολουθεί.

Η χρησιμοποίηση της πλεγματικής μορφής (lattice), οφείλεται στην δυνατότητα που δίνουν οι συντελεστές ανάκλασης για παραμετρική παρεμβολή (interpolation), μεταξύ φίλτρων που αντιστοιχούν σε διαφορετικά τμήματα φωνής. Η παρεμβολή, δηλαδή, μεταξύ ευσταθών φίλτρων, μέσω της παρεμβολής των συντελεστών ανάκλασης k_i , για τους οποίους ισχύει $-1 < k_i < 1$, θα δώσει πάλι ευσταθές φίλτρο ($|k_i| < 1$). Η δυνατότητα αυτή της πλεγματικής μορφής και της θεωρίας Γραμμικής Πρόβλεψης γενικότερα, καθιστούν ιδιαίτερα αποτελεσματική την φασματική εξομάλυνση μεταξύ ανόμοιων τμημάτων φωνής σε τεχνικές αλυσιδωτής σύνδεσης (concatenation synthesis).



[Η πλεγματική μορφή που χρησιμοποιείται στην Σύνθεση Φωνής. Η είσοδος είναι το συνθετικό σήμα διέγερσης (σήμα σφάλματος $e(n)$) και έξοδος το παραγόμενο σήμα φωνής $s(n)$]

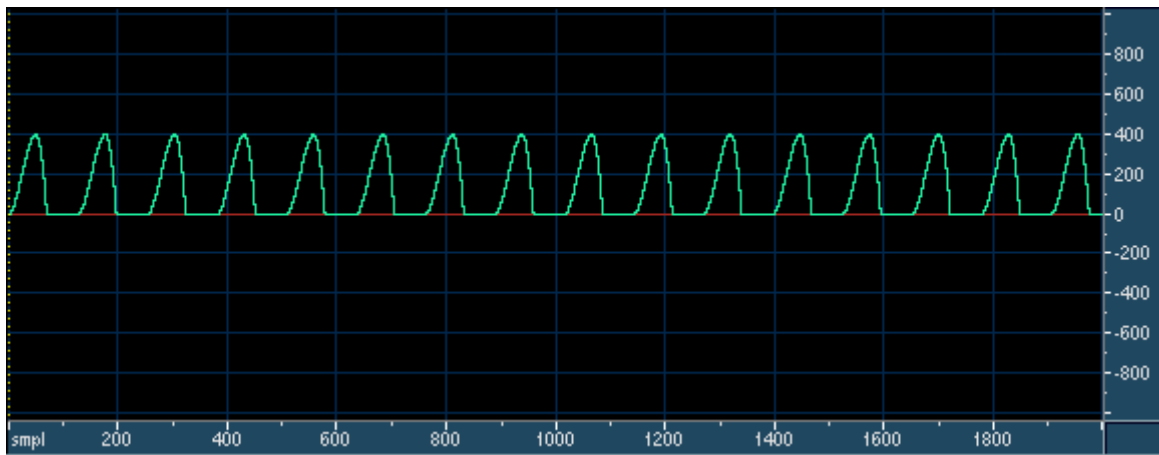
Στην εργασία αυτή, οι συντελεστές k_i , βρέθηκαν μέσω του υπολογισμού των συντελεστών πρόβλεψης a_i . Στην πράξη, ο υπολογισμός των συντελεστών k_i γίνεται πιο άμεσα με την χρησιμοποίηση, πάλι πλεγματικής μορφής. Οι συντελεστές ανάκλασης k_i ονομάζονται σ' αυτήν την περίπτωση PARCOR (PARTIAL CORrelation) συντελεστές, εξαιτίας του ότι υπολογίζονται από την συσχέτιση (correlation) μεταξύ των σφαλμάτων πρόβλεψης (forward & backward) και αποδεικνύεται πως είναι ίσοι με τους συντελεστές k_i που προκύπτουν από την μέθοδο του Durbin, όπως αναλύθηκε παραπάνω. Για περισσότερες πληροφορίες, ο ενδιαφερόμενος μπορεί να αναφερθεί στην σχετική βιβλιογραφία (Markel & Gray, 1976, Rabiner & Schafer 1978).

Η παραπάνω πλεγματική υλοποίηση, όπως είπαμε, μετατρέπει την φασματική εξομάλυνση (spectral smoothing), σε εύκολο interpolation των συντελεστών ανάκλασης. Όμως, η πλεγματική υλοποίηση είναι δυνατή μόνο εάν το φίλτρο $H(z)$, είναι ένα all-pole φίλτρο, δηλαδή μόνο εάν ο φωνητικός σωλήνας προσομοιωθεί με ένα ψηφιακό φίλτρο που έχει μόνο πόλους και όχι μηδενικά. Η υπόθεση αυτή, είναι φτωχή στην περίπτωση των ρινικών φθόγγων (**nasals**), εξαιτίας της σύζευξης της ρινικής με την στοματική κοιλότητα. Η απαίτηση μηδενικών είναι, επίσης έκδηλη, στην περίπτωση των προστρίβόμενων φθόγγων (**fricatives**), αλλά αντιμετωπίζεται με αύξηση της τάξης p του φίλτρου $H(z)$ (και επομένως του αριθμού των συντελεστών πρόβλεψης).

Ιδιαίτερη προσοχή πρέπει να δοθεί στους κροτικούς φθόγγους (**plossives**), εξαιτίας της μικρής διάρκειας του διαστήματος έκρηξης, πολύ μικρότερης, ενδεχομένως από την διάρκεια του διαστήματος ανάλυσης (συνήθως 30 ms) και συγκρινόμενης της διάρκειας του διαστήματος μετατόπισης (time shift $\approx 6-10$ ms) του πλαισίου ανάλυσης.

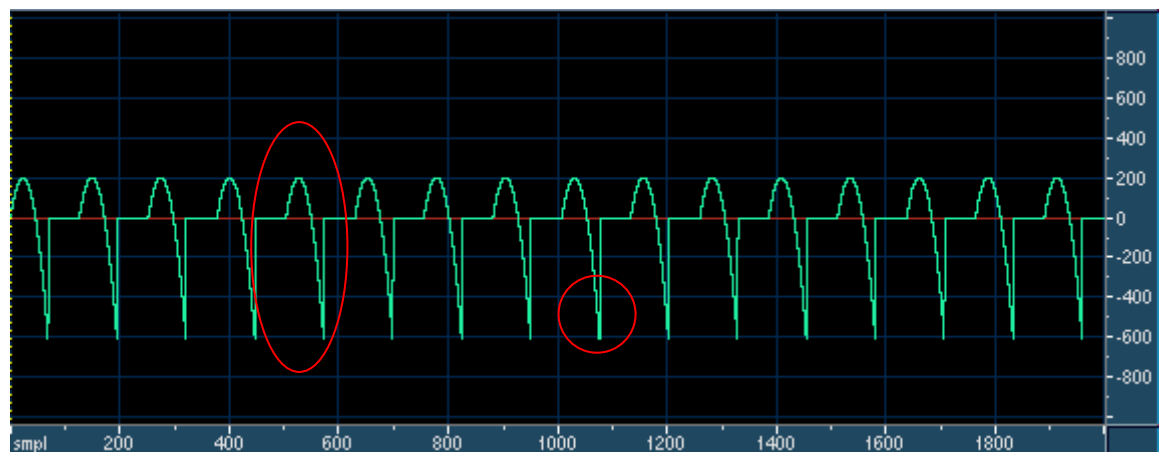
Από τα παραπάνω, είναι προφανές ότι η θεωρία LPC μετατρέπει το πρόβλημα Σύνθεσης Φωνής σε πρόβλημα κατασκευής σήματος διέγερσης $e(n)$ στο φίλτρο $H(z)$ του φωνητικού σωλήνα, τόσο για τους έμφωνους (voiced) όσο και για τους άφωνους (unvoiced) φθόγγους.

Οι πρώτες προσεγγίσεις στο παραπάνω θέμα, χρησιμοποίησαν σειρές παλμών με περίοδο όσο το επιθυμητό pitch, για τους έμφωνους φθόγγους και λευκό θόρυβο για τους άφωνους φθόγγους. Χαρακτηριστικό μοντέλο αυτής της κατηγορίας, είναι το μοντέλο του Rosenberg για τους έμφωνους φθόγγους (Rosenberg, 1971), στο οποίο επιχειρείται η μίμηση των γλωτιδικών παλμών, όπως εμφανίζονται στο πεδίο του χρόνου. Συγκεκριμένα, χρησιμοποιείται πολώνυμο 3ης τάξης για το άνοιγμα των φωνητικών χορδών και πολώνυμο 2ης τάξης για το κλείσιμο των φωνητικών χορδών (glottal opening & closure).



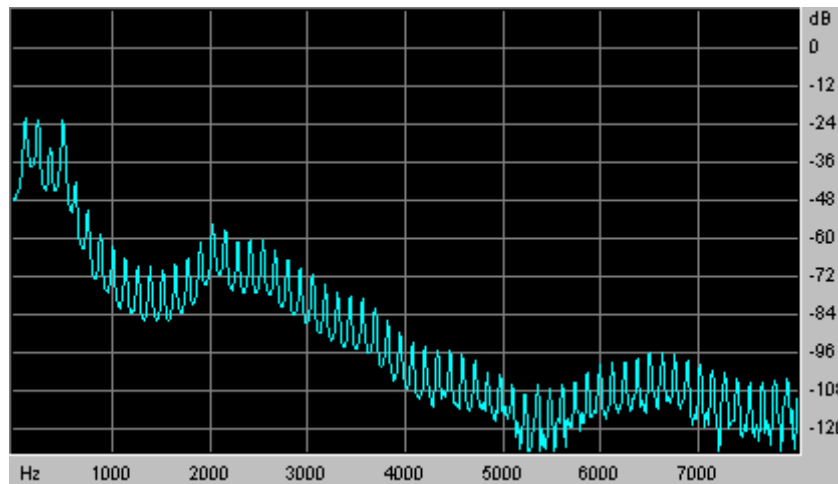
[το σήμα διέγερσης $e(n)$ για τους έμφωνους φθόγγους, όπως προτάθηκε από τον Rosenberg]

Βελτίωση του παραπάνω μοντέλου αποτελεί η προσθήκη από τον Oliveira (Oliveira, 1993-94) ενός εκθετικού παράγοντα που ελλατώνεται, κατά την φάση που η γλωττίδα κλείνει.



[το μοντέλο του Rosenberg για τους έμφωνους φθόγγους, τροποποιημένο από τον Oliveira. Η προσθήκη του εκθετικού, με το “απότομο” σημείο καμπής, προσθέτει το απαραίτητο υψίσυχο περιεχόμενο που κάθε σήμα διέγερσης πρέπει να έχει]

Τα παραπάνω απλοποιημένα μοντέλα διέγερσης, κάθε άλλο παρά ικανοποιητικά αποτελέσματα σύνθεσης δίνουν. Το σήμα σφάλματος-διέγερση, περιέχει επιπλέον πληροφορία που “διορθώνει” την παραδοχή ότι όλοι οι φθόγγοι μπορούν να παραχθούν από ένα all-pole φίλτρο. Μοντέλα διέγερσης που δεν περιέχουν αυτήν την επιπλέον πληροφορία αποτυγχάνουν να συνθέσουν “φυσικά”. Η “φυσικότητα” της φωνής που παράγεται είναι μειωμένη καθώς οι παραγόμενοι φθόγγοι υποφέρουν από χαρακτηριστικό “μπαζ” (LPC buzziness).



[Μετασχηματισμός Fourier συνθετικού /e/, με το μοντέλο σύνθεσης του Rosenberg. Φαίνεται καθαρά η απουσία υψίσυχνου “θορύβου”, ενώ είναι εμφανής η παρουσία του pitch και των αρμονικών του. Ο φθόγγος “υποφέρει” από buzziness]

Προς την κατεύθυνση εισαγωγής της “επιπλέον πληροφορίας” στα μοντέλα διέγερσης, έχουν επικεντρωθεί οι ερευνητικές προσπάθειες τα τελευταία χρόνια. Χαρακτηριστική, είναι και πάλι η προσπάθεια του Oliveira (Oliveira 1993-94), ο οποίος κατόρθωσε να ενσωματώσει στο τροποποιημένο μοντέλο του Rosenberg, συντελεστές που εκφράσουν “aspiration noise” τους οποίους υπολόγισε με τεχνική Least Squares, από το αρχικό, “φυσικό” σήμα. Το μοντέλο του, που επιπλέον ενσωματώνει περίπλοκες τεχνικές εξομάλυνσης μεταβάσεων μεταξύ έμφωνων και άφωνων τμημάτων φωνής, χρησιμοποιείται αυτή την στιγμή, στο text-to-speech system της AT&T, με επιτυχία.

Λύση στο πρόβλημα των φτωχών μοντέλων διέγερσης, αποτελεί η εφαρμογή MBROLA τεχνικής στο εναπομένον σήμα διέγερσης, καταλήγοντας σε μια υβριδική LP-MBROLA τεχνική!

Επίσης, γρήγορη λύση θα έδινε η εισαγωγή μηδενικών ή η εξαγωγή δειγμάτων από το εναπομένον σήμα διέγερσης (residual), για την μεταβολή της βασικής συχνότητας.

Συμπέρασμα:

Η επιτυχία επομένως, μιας LP τεχνικής για σύνθεση φωνής σκοντάφτει στην επιλογή ενός ικανού μοντέλου διέγερσης.

Επιπλέον, η φιλοσοφία της θεωρίας γραμμικής πρόβλεψης υιοθετεί την ανεξαρτησία του φωνητικού σωλήνα (vocal tract) από την διέγερση, παραδοχή που όμως δεν επιβεβαιώνεται

από την πράξη. Ο Klatt (Klatt, 1987) αναφέρει πως οι συντονισμοί (resonances) στον φωνητικό σωλήνα οδηγούν σε στάσιμα κύματα πίεσης που επηρεάζουν την ταλάντωση των φωνητικών χορδών και επομένως το σχήμα των γλωττιδικών παλμών. Αντιστρόφως, το ανοιγοκλείσιμο των φωνητικών χορδών αντιστοιχεί σε μια μεταβαλλόμενη σύνθετη αντίσταση τερματισμού για τον φωνητικό σωλήνα, η οποία μεταβάλλει την χαρακτηριστική συνάρτηση μεταφοράς (transfer function) του vocal tract.

Συνεπώς, οι παραπάνω δυσκολίες, πρέπει να συνυπολογιστούν κατά την αξιολόγηση και εφαρμογή της θεωρίας Γραμμικής Πρόβλεψης στην Σύνθεση Φωνής.

Κεφάλαιο τρίτο

Παραγωγή Ακουστικών-Φωνητικών Μονάδων

Όπως τονίστηκε στο προηγούμενο κεφάλαιο, η τεχνική σύνθεσης που εφαρμόστηκε στον Αίσωπο, είναι μια τεχνική αλυσιδωτής σύνδεσης *προηχογραφημένων τμημάτων φωνής* (concatenation synthesis). Συνεπώς, αυτά τα τμήματα φωνής αποτελούν απαραίτητες μονάδες για τον Συνθέτη φωνής του συστήματος.

Στην ενότητα αυτή, θα σχολιάσουμε την επιλογή του είδους των απαραίτητων φωνητικών-ακουστικών μονάδων, θα περιγράψουμε το είδος των φθόγγων της Ελληνικής Γλώσσας, θα ορίσουμε τον συμβολισμό τους και τέλος, θα παραθέσουμε την μεθοδολογία, βάση της οποίας κατασκευάσαμε τις φωνητικές μονάδες.

3.1 Είδος Φωνητικών Μονάδων - Γιατί Δίφωνα;

Τα τμήματα φωνής, που χρησιμοποιούνται ως μονάδες σύνθεσης οφείλουν να έχουν διάρκεια, όσο το δυνατόν μεγαλύτερη, έτσι ώστε να διατηρούν την φυσικότητα της φωνής σε μεγαλύτερα διαστήματα χρόνου, απαιτώντας ταυτόχρονα λιγότερα σημεία ένωσης (concatenation points) κατά την αλυσιδωτή σύνδεση. Από την άλλη μεριά, ο αριθμός των φωνητικών μονάδων πρέπει να διατηρείται σε λογικά επίπεδα για πιο εύκολο χειρισμό τους από το text-to-speech σύστημα.

Η πρώτη προσέγγιση στο θέμα από κάποιον όχι σχετικό με το αντικείμενο, θα ήταν η χρησιμοποίηση ολόκληρων λέξεων ως φωνητικές μονάδες. Πράγματι, με την χρησιμοποίηση λέξεων επιτυγχάνεται ο εγκλωβισμός “δύσκολων” φαινομένων συνάρθρωσης (**coarticulation effects**¹), εντός των λέξεων.

Ωστόσο, η προφορά απομονωμένων λέξεων, διαφέρει από την προφορά των ίδιων λέξεων μέσα σε φράσεις, καθώς τα προσωδικά τους χαρακτηριστικά εξαρτώνται από το περιεχόμενο των φράσεων στις οποίες ανήκουν. Η διάρκεια, για παράδειγμα, των απομονωμένων λέξεων είναι μεγαλύτερη από την διάρκεια των ίδιων λέξεων σε φράσεις. Η εισαγωγή τεχνικών ψηφιακής επεξεργασίας για τον έλεγχο της προσωδίας σε επίπεδο λέξης, είναι πολύ δύσκολη, αν όχι αδύνατη λόγω της πληθώρας μεταβάσεων από φθόγγο σε φθόγγο μέσα στη “λέξη” τις οποίες θα κληθούν να αντιμετωπίσουν.

¹ Η συνάρθρωση, ουσιαστικά, υπογραμμίζει τον επηρεασμό της άρθρωσης κάθε φθόγγου από τους γειτονικούς του. Το πώς αρθρώνεται (=σχηματίζεται) ένας φθόγγος εξαρτάται όχι μόνο από το είδος του φθόγγου, αλλά και από την “γειτονιά” του (phonetic context)

Επιπλέον, δεν θα πρέπει να ξεχνάμε πως η λέξη είναι μια γλωσσική μονάδα που προσδιορίζει μια έννοια. Όταν μιλάμε, μεταξύ δύο σημαντικών παύσεων προφέρουμε πλήθος λέξεων, σχηματίζοντας “εκφωνήσεις”. Συνεπώς, σε φωνητικό επίπεδο, δεν υπάρχουν λέξεις, αλλά τον ρόλο των αυτόνομων δομικών μονάδων τον έχουν οι “εκφωνήσεις” (Βαλεοντής & Καγιάφας, 1984).

Ακόμη και εάν ηχογραφήσαμε πλήθος των ίδιων λέξεων σε διαφορετικά ακουστικά περιβάλλοντα, θα χρειαζόμασταν τρομακτικό αριθμό λέξεων, γεγονός όχι αμελητέο.

Πέρα, όμως από τα παραπάνω, η Σύνθεση Φωνής με χρήση λέξεων, εκ των πραγμάτων οδηγεί σε Φωνή περιορισμένου λεξιλογίου.

Η χρήση συλλαβών, ως φωνητικές μονάδες, θα μπορούσε να περιορίσει τον συνολικό απαιτούμενο αριθμό μονάδων, ωστόσο η χρήση τους οδηγεί σε φασματικές ασυνέχειες στα σημεία ένωσης, εξαιτίας έντονων φαινομένων συνάρθρωσης μεταξύ των συλλαβών. Για τον ίδιο λόγο, και η χρήση ημισυλλαβών, όπως στο εμπορικό σύστημα της Bellcore (ORATOR TTS), οδηγεί σε ακουστές ασυνέχειες, παρά τις όποιες τεχνικές εξομάλυνσης (π.χ. με χρήση θεωρίας γραμμικής πρόβλεψης).

Στον Αίσωπο, χρησιμοποιήθηκαν τα δίφωνα, τμήματα φωνής που περιλαμβάνουν steady-state φθόγγο (phone), μετάβαση στον επόμενο (transition) και steady-state του επόμενου φθόγγου. Περιλαμβάνουν, δηλαδή μεταβάσεις φθόγγων που είναι δύσκολο να παραχθούν. Επιπλέον, επειδή τα φασματικά χαρακτηριστικά steady-state φθόγγων (phones) είναι, σχετικώς, σταθερά στις διάφορες φωνητικές “γειτονιές” (phonetic contexts), οι επιδράσεις της συνάρθρωσης στα σημεία ένωσης ελαχιστοποιούνται, παράγοντας σχετικώς “ομαλές” συνδέσεις μονάδων σύνθεσης σε μια concatenation υλοποίηση.

Για τους παραπάνω λόγους, το δίφωνο, είναι το πιο διαδεδομένο είδος μονάδας σύνθεσης που αυτή την στιγμή χρησιμοποιείται, στους πιο πετυχημένους Συνθέτες Φωνής.

Ο απαιτούμενος αριθμός τους δε, κυμαίνεται σε ‘υποφερτά’ επίπεδα:

Γλώσσα	Αριθμός διφώνων
Αγγλικά	1027
Ισπανικά	501
Γερμανικά	1102
Γαλλικά	1123
Ιαπωνικά	445
Ελληνικά	

(ESoPoS-TTS)

540

[αριθμός απαιτούμενων διφώνων, για τις Γλώσσες που μελετήθηκαν (εκτός από τα Ελληνικά) στα εργαστήρια της Bell (The Bell Labs Approach, 1998)]

3.2 Συμβολισμός Φωνητικών μονάδων

Αφού καταλήξαμε στο δίφωνο, ως το είδος των φωνητικών μονάδων που θα χρησιμοποιήσουμε, και αφού το ορίσαμε ως το τμήμα φωνής που περιλαμβάνει steady-state φθόγγο, μετάβαση στον επόμενο φθόγγο και steady-state του επόμενου φθόγγου, οφείλουμε να μελετήσουμε και να συμβολίσουμε τους φθόγγους (phones) από τους οποίους θα σχηματιστούν τα δίφωνα.

3.2.1 Βασικές Έννοιες

Προτού όμως προχωρήσουμε, θα πρέπει να ξεκαθαρίσουμε ορισμένες βασικές έννοιες που κατα κόρον θα χρησιμοποιηθούν στην συνέχεια:

- **Φώνημα (phoneme):**

Είναι η ελάχιστη γλωσσική μονάδα με εννοιολογικά διαφοροποιητική αξία. Η αντικατάσταση ενός φωνήματος από ένα άλλο, στο ίδιο φωνητικό περιβάλλον επιφέρει αλλαγή σημασίας π.χ /p/ -> /f/ στην λέξη /ponos/ -> /fonos/

- **Φθόγγος (phone):**

Η απλή φωνή που αποτελεί συστατικό της ηχητικής υπόστασης της γλώσσας. Δεν θα πρέπει να συγχέεται με το φώνημα, καθώς ο φθόγγος (phone) αποτελεί την ακουστική έκφραση του φωνήματος (phoneme), ενώ το φώνημα αποτελεί τον γραπτό συμβολισμό του φθόγγου. Μάλιστα, δύο ή περισσότεροι παραπλήσιοι φθόγγοι μπορούν να αντιπροσωπεύονται από το ίδιο φώνημα π.χ το /a/ μπορεί να προφερθεί στοματικά ή ρινικά, σε κάθε περίπτωση όμως, αντιστοιχεί στο φώνημα /a/.

- **Γράμμα:**

Είναι το ελάχιστο ξεχωριστό μέρος του συμβόλου ενός φωνήματος π.χ. στην λέξη κάτοικος, τα γράμματα <o>, <i> συμβολίζουν το φώνημα /i/

Είναι προφανές ότι δεν υπάρχει αμφιμονοσήμαντη σχέση μεταξύ γραμμάτων και φωνημάτων.

- **Λέξη:**

Η γλωσσική μονάδα που χρησιμοποιούμε για να συμβολίσουμε μια έννοια, και της οποίας η γραπτή μορφή μέσα σε ένα κείμενο βρίσκεται ανάμεσα σε δύο κενά διαστήματα.

- **Εκφώνηση:**

Είναι μέρος “προφορικού” κειμένου που εκφέρεται χωρίς καμμία ενδιάμεση σημαντική παύση π.χ στην φράση “καλημέρα σας, τί κάνετε;” διακρίνουμε δύο εκφωνήσεις: /kalimerasas/ + /tikanete/

3.2.2 Κατάταξη των φθόγγων της Ελληνικής

Οι φθόγγοι της Ελληνικής χωρίζονται σε φωνήεντα και σύμφωνα.

Οι φθόγγοι-φωνήεντα είναι 5: α, ε, ι, ο, ου, ενώ

οι φθόγγοι-σύμφωνα είναι 20 + 8: β, γ, δ, ζ, θ, κ, λ, μ, ν, π, ρ, σ, τ, φ, χ, μπ, ντ, γκ, τσ, τζ + τις ουρανικές παραλλαγές μερικών: γ', κ', λ', μ', ν', ν'', χ', γκ'

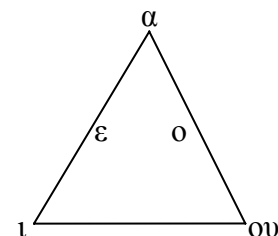
π.χ. γ'έρος και γωνία, κ'αιρός και καλός, λ'ιάζομαι και λουλούδι, ν'οιάζομαι και νερό, άν'γκ'ελος και γκρεμίζω, χ'έρι και χωνί.

[παραστατικότερος φωνηματικός συμβολισμός των φθόγγων και παραδείγματα, θα δοθούν παρακάτω]

- **φωνήεντα**

Χωρίζονται σε ουρανικά (/ε/, /ι/) και σε υπερωιακά (/ο/, /ου/). Το /α/ στο οποίο το στόμα ανοίγει περισσότερο, βρίσκεται στην κορυφή του τριγώνου, ενώ στις δύο άλλες κορυφές βρίσκονται το /ι/ και το /ου/.

Ενδιάμεσα, βρίσκεται το /ε/ όταν το στόμα κλείνει από το /α/ προς το /ι/ και το /ο/, όταν το στόμα κλείνει από το /α/ προς το /ου/.



[το τρίγωνο των φωνηέντων φθόγγων]

- **σύμφωνα**

Ο παρακάτω πίνακας είναι διαφωτιστικός:

Κατά τα μέρη του φωνητικού σωλήνα όπου σχηματίζονται		Κατά τη διάρκεια			
		Στιγμιαία		Εξακολουθητικά	
		Άηχα	Ηχηρά	Τριβόμενα Άηχα	Ρινικά Υγρά
Χειλικά	Διχειλικά	/π/	/μπ/		/μ/ /μ'/
	Χειλοδοντικά			/φ/	/β/
	Μεσοδοντικά			/θ/	/δ/
Οδοντικά	Γλωσσοδοντικά	/τ/	/ντ/		
	Διπλοδοντικά	/τσ/	/τζ/	/σ/	/ζ/

Λαρυγγικά	Ουρανικά	/κ'/	/γκ'/	/χ'/	/γ'/	/ν''(γκ)/
	Υπερωικά	/κ/	/γκ/	/χ/	/γ/	
Γλωσσικά	Ουρανικά				/ν'/	/λ'/
	Υπερωικά				/ν/	/λ/ /ρ/

[Ο πίνακας των συμφώνων (Τριανταφυλλίδης, 1941-96)]

Ιδιαίτερη μνεία θα πρέπει να γίνει στα μισόφωνα, τα οποία είναι ενδιάμεσα φωνηέντων και συμφώνων. Τα μισόφωνα της Ελληνικής, είναι το /ι/ και το /ου/, όταν προφέρονται όπως στο 'πιάνο' και στο 'ουίσκι'.

Επίσης θα πρέπει να τονίσουμε την ύπαρξη καταχρηστικών διφθόγγων, στους οποίους το προταχτικό /ι/, έχει την φωνητική αξία μισοφώνου. Στους διφθόγγους αυτούς, το /ι/ ακολουθείται από άλλο φωνήεν, αποτελώντας έναν δίφθογγο στον οποίο η προφορά του μισοφώνου /ι/ εξαρτάται από τον φθόγγο που προηγείται²:

α) προφέρεται σαν /γ'/, όταν προηγείται ηχηρό /μπ/, /ντ/, /β/, /ζ/, /δ/, /ρ/ π.χ. κάμπια, κατάντια, βιάζομαι κλπ.

β) προφέρεται, συνήθως, σαν /χ'/ όταν προηγείται άηχο /π/, /τ/, /φ/, /θ/, /σ/ π.χ. ποιές, ματιές, καρφιά, ίσιωμα κλπ.

γ) ύστερα από λ, ν, κ, γ, γκ, χ δεν προφέρεται καθόλου, αλλά χρησιμεύει ως *διακριτικό γράμμα* για να δείξει πως τα παραπάνω προφέρονται ουρανικά (/λ'/, /ν'/, /κ'/, /γ'/, /γκ', /χ'/) και όχι υπερωικά (/λ/, /ν/, /κ/, /γ/, /γκ/, /χ/) π.χ. ελ'ιά, ενν'ιά, γ'ιατί και όχι γατί κλπ.

Οι καταχρηστικοί δίφθογγοι, είναι υπεύθυνοι για το φαινόμενο της συνίζησης, το οποίο δυσκολεύει την φωνηματική μεταγραφή των λέξεων και για το οποίο θα μιλήσουμε παρακάτω.

3.2.3 Το Φωνηματικό Αλφάβητο του "Αίσωπον"

Αφού κατηγοριοποιήσαμε τους φθόγγους της Ελληνικής (phones), συμβολίζοντάς τους με απλό εμπειρικό τρόπο (phonemes), θα πρέπει να καταλήξουμε σε έναν συμβολισμό των

φθόγγων, εύκολο στον χειρισμό και συμβατό με το σύνολο των συμβόλων που ο H/Y μπορεί να χειριστεί. Ας μην ξεχνάμε, πως πρέπει με κάποιο τρόπο να συμβολίσουμε όλους τους συνδυασμούς των φθόγγων (δίφωνα).

Ουσιαστικά, αναζητούμε έναν μετασχηματισμό των συμβόλων του Διεθνούς Φωνητικού Αλφαβήτου (International Phonetic Alphabet-I.P.A.) σε 7-bit ASCII σύμβολα. [Το I.P.A. δεν μπορεί να χρησιμοποιηθεί απευθείας, αφού περιέχει μίξη λατινικών και ελληνικών χαρακτήρων]

Ευτυχώς, ο μετασχηματισμός αυτός (IPA → ASCII), έγινε από εξειδικευμένη ομάδα επιστημόνων, για τις περισσότερες Ευρωπαϊκές Γλώσσες, στα πλαίσια του Ευρωπαϊκού προγράμματος ESPRIT (project 1541).

Το αποτέλεσμα, είναι το αλφάβητο S.A.M.P.A.³ (Speech Assesment Methods Phonetic Alphabet) που για τα Ελληνικά, παρουσιάζεται παρακάτω:

- **Φωνήεντα (Vowels)**

i, e, a, o, u

- **Σύμφωνα (Consonants)**

Κροτικά -Στιγμαία (plosives)

Σύμβολο. Glossary (SAMPA) Ορθογραφία

p	I drink	"pino	πίνω
b	I can	bo"ro	μπορώ
t	then	"tote	τότε
d	I dress	"dino	ντύνω
k	I do	"kano	κάνω
g	I throw down	gre"mizo	γκρεμίζω

Διπλά (affricates)

ts	pocket	"tsepi	τσέπη
dz	freeloader	dzaba"dzis	τσαμπατζής

Προστριβώμενα-Εξακολουθητικά (fricatives)

f	I depart	"fevGo	φεύγω
v	evening	"vraDi	βράδυ

² το /ɪ/ στους καταχρηστικούς δίφθογγους (κατά SAMPA συμβολίζεται με /j/), αποτελεί το φώνημα (phoneme) ενώ η ακουστική του αντιπροσώπευση ανάλογα με τον φθόγγο που προηγείται, αποτελεί το phone (φθόγγο)

³ για περισσότερες πληροφορίες σχετικά με το S.A.M.P.A. → <http://www.phon.ucl.ac.uk/home/sampa>

T	I want	"Telo	θέλω
D	route	"Dromos	δρόμος
s	house	"spiti	σπίτι
z	heat	"zesti	ζέστη
x	time	"xronos	χρόνος
G	milk	"Gala	γάλα

Ρινικά (nasals)

m	I speak	mi"lo	μιλώ
n	water	ne"ro	νερό
N	cucumber	an"guri	αγγούρι

([N] = αλλόφωνο του /n/)

Υγρά (liquids)

l	flower	lu"luDi	λουλούδι
r	clothes	"ruxa	ρούχα

Μισόφωνα (semivowel)

j	I read	Dja"vazo	διαβάζω
---	--------	----------	---------

Ουρανικά (palatals)

c	and	ce	και
	I sleep	ci"mame	κοιμάμαι
	I frighten	"scazo	σκιάζω
gj	angel	"angjelos	άγγελος
	bad luck	"gjinja	γκίνια
	screech-owl	"gjonis	γκιώνης
C	hand	"Ceri	χέρι
	winter	Ci"monas	χειμώνας
	snow	"Coni	χιόνι
jj	old man	"jjeros	γέρος
	circle	"jjiros	γύρος
	son	jjos	γιός

3.2.4 Καταγραφή Φωνητικών-Ακουστικών Μονάδων

Τα 31 φωνηματικά σύμβολα του προγράμματος SAMPA, ενσωματώνουν τους διαχωρισμούς των φθόγγων, όπως τους περιγράψαμε στην ενότητα 3.2.2. Το φωνηματικό αλφάβητο που προκύπτει, είναι μια απλοποιημένη (όχι απλοϊκή) έκδοση του Διεθνούς Φωνητικού Αλφαβήτου, προσαρμοσμένη στις ιδιαιτερότητες μιας ψηφιακής text-to-speech εφαρμογής.

Θεωρητικά τα 31 φωνήματα συν την παύση (που συμβολίζεται με _), δίνουν $32^2=1024$ πιθανά δίφωνα. Στο σημείο αυτό θα πρέπει να αναφέρουμε πως στο σύνολο των δίφωνων, πρέπει να συμπεριληφθούν και τα δίφωνα που εμφανίζονται μεταξύ των λέξεων. Για παράδειγμα, ενώ δεν υπάρχει λέξη που να περιέχει το δίφωνο /kc/, το δίφωνο αυτό μπορεί να εμφανιστεί σε μια φράση όπως “Πάοκ και Άρης” → /paok ce aris/.

Επιπλέον, θα πρέπει να λάβουμε υπόψιν τα απαραίτητα δίφωνα, για την εκφώνηση ξένων λέξεων που χρησιμοποιούνται στην Ελληνική. Για παράδειγμα, η λέξη γκάλοπ (=δημοσκόπηση) τελειώνει στον φθόγγο /p/, συνεπώς θα πρέπει να συμπεριληφθούν όλα τα δίφωνα μεταξύ λέξεων εκ των οποίων η πρώτη τελειώνει σε /p/.

Οι παραπάνω παρατηρήσεις λήφθηκαν υπόψιν, κατά την καταγραφή των δίφωνων. Ωστόσο, **φωνοτακτικοί** περιορισμοί περιόρισαν αισθητά το πλήθος των απαιτούμενων δίφωνων: ορισμένα από τα 31 φωνήματα αποτελούν αλλοφωνικές εκδόσεις κάποιων φωνημάτων, συνεπώς κάποιοι συνδυασμοί φωνημάτων a priori απαλείφονται. Π.χ. το /C/ όπως στο “χέρι” (/Ceri/), αποτελεί την ουρανική έκδοση του /x/ (υπερωικό) όπως στο “χωνί” (/xoni/), συνεπώς συνδυασμός /xe/ δεν υφίσταται (υφίσταται μόνο /ce/).

Επιπλέον, στα παραπάνω φωνήματα, συμπεριλάβαμε το μισόφωνο /j/ , το οποίο εμφανίζεται σε συγκεκριμένους συνδυασμούς συμφώνων με καταχρηστικούς δίφθογγους, συνεπώς ο αριθμός των δίφωνων στα οποία συμμετέχει είναι περιορισμένος.

Καταλήξαμε, σε ένα αριθμό 540 απαραίτητων δίφωνων, και για κάθε δίφωνο αναζητήσαμε, τουλάχιστον 2 λέξεις ή φράσεις, στις οποίες το δίφωνο αυτό εμφανίζεται παρακάτω.

Ξεκαθαρίστηκε από την αρχή, πως τα δίφωνα θα εξαχθούν μέσα από φράσεις Προφορικού Λόγου, οι οποίες θα εκφωνηθούν με “φυσικό” ρυθμό και ουδέτερο “χρωματισμό”(=σταθερό pitch). Στην βιβλιογραφία, έχει αναφερθεί η εξαγωγή των μονάδων από συνδυασμούς γραμμάτων χωρίς νόημα (logatoms, Dutoit 1997). Ωστόσο, σ’ αυτήν την περίπτωση, είναι **δύσκολο** για τον **εκφωνητή η διατήρηση ενός σταθερού ρυθμού ομιλίας**, με συνέπεια μέσες τιμές διάρκειας φθόγγων, που απέχουν από τις διάρκειες που χρησιμοποιούμε όταν μιλάμε. Η εκφώνηση με φυσικό ρυθμό των λέξεων

απ'όπου θα εξαχθούν τα δίφωνα, θα δώσει στην συνέχεια την δυνατότητα, για μια **αξιόπιστη στατιστική αποτίμηση της διάρκειας του κάθε φθόγγου.**

Επιπλέον, προτιμήθηκαν λέξεις στις οποίες τα δίφωνα δεν τονίζονται, έτσι ώστε να υποβοηθηθεί περισσότερο, μια πιο ουδέτερη εκφώνηση.

Οι λέξεις και οι φράσεις εντός των οποίων υπάρχουν τα ζητούμενα δίφωνα, επιλέχθηκαν με συγκεκριμένο τρόπο, έτσι ώστε τα δίφωνα να βρίσκονται σε γειτονιές ελάχιστης συνάρθρωσης (minimum coarticulation). Οι συνδυασμοί συμφώνων φθόγγων, όπου ο καθένας επηρεάζει το τρόπο σχηματισμού του επόμενου όσο το δυνατόν λιγότερο, φαίνεται παρακάτω:

αρχικό σύμφωνο	επόμενο σύμφωνο
στιγμαίο	στιγμαίο
στιγμαίο	ρινικό
τριβικό	τριβικό
τριβικό	ρινικό
τριβικό	στιγμαίο
ρινικό	στιγμαίο
ρινικό	τριβικό
υγρό	στιγμαίο
υγρό	τριβικό

[ζευγάρια συμφώνων ελάχιστης συνάρθρωσης (minimum coarticulation) (The Bell Labs Approach, 1998)]

Για παράδειγμα, για την εξαγωγή του διφώνου /no/ θα προτιμηθεί η λέξη “άυπνος” (/aip**nos**/) από τη λέξη “ψέλνω” (/psel**no**/) αφού στην πρώτη, το ζητούμενο δίφωνο “γειτονεύει” με το /p/ (στιγμαίο) και επομένως σχηματίζεται το ζευγάρι ελάχιστης συνάρθρωσης /pn/, ενώ στην δεύτερη λέξη το δίφωνο γειτονεύει με το σύμφωνο /l/ (υγρό) και ο συνδυασμός /ln/, δεν είναι ζευγάρι ελάχιστης συνάρθρωσης.

Για την επιλογή των λέξεων, χρησιμοποιήθηκε λεξικό σε ηλεκτρονική μορφή [“νέο Λεξικό της Ελληνικής- MLS-Εκδόσεις Σταφυλίδης”].

3.3 Εξαγωγή Διφώνων

Αφού ολοκληρώθηκε ο συμβολισμός και η καταγραφή των διφώνων, ηχογραφήθηκε κάθε λέξη που περιλαμβάνει το δίφωνο μέσα στην φράση “πες την λέξη <λέξη> ξανά”, έτσι ώστε να διατηρηθεί ένας σταθερός και φυσικός ρυθμός ομιλίας ενώ παράλληλα να υποβοηθηθεί ο εκφωνητής στην διατήρηση ενός σταθερού-ουδέτερου pitch. Η παραπάνω φράση, εκφωνήθηκε ουδέτερα για όλες τις λέξεις (και επομένως για όλα τα δίφωνα) σε μια περίοδο ηχογράφησης, έτσι ώστε να διατηρηθούν σταθερές οι συνθήκες ηχογράφησης (π.χ. μέση ενέργεια, ανά κατηγορία φθόγγων, βασική συχνότητα F_0).

Οι ηχογραφήσεις έγιναν στο Εργαστήριο Ηλεκτρακουστικής του τμήματος Ηλεκτρολόγων Μηχανικών, στα 44.1 kHz με 16 bit/sample και με ποιότητα ηχογράφησης CD μουσικής (CD quality). Τα δείγματα, στην συνέχεια με υποδειγματοληψία “κατέβηκαν” στα 16 kHz. Στην συνέχεια, έγινε ο έλεγχος της βασικής συχνότητας (pitch) για το κάθε ηχογραφημένο δίφωνο, με την μέθοδο Average Magnitude Differential Function (A.M.D.F.)⁴

Ο εκφωνητής μας, διατήρησε σταθερή την βασική του συχνότητα κοντά στα 100 Hz, για το 90% των διφώνων με μια απόκλιση περίπου + 5%. Για τα δίφωνα με pitch εκτός του ορίου του +5%, αναζητήθηκαν άλλες λέξεις στο σύνολο των ηχογραφήσεων, που περιείχαν τα ζητούμενα δίφωνα με pitch κοντά στα 100 Hz, ενώ για όσα δίφωνα αυτό δεν ήταν δυνατόν, έγινε συμπληρωματική ηχογράφηση.

Αν και με την MBROLA τεχνική, τα δίφωνα θα ανασυνθεθούν, έτσι ώστε να επιτευχθεί σταθερό pitch όπως περιγράψαμε σε προηγούμενο κεφάλαιο, είναι σημαντικό οι έμφωνοι φθόγγοι των διφώνων, να έχουν εξαρχής σταθερή βασική συχνότητα. Όσο λιγότερο σταθερή είναι η βασική συχνότητα των διφώνων, τόσο πιο απαραίτητη θα είναι η ψηφιακή επεξεργασία τους, και επομένως τόσο πιο μεγάλη η αλλοίωσή τους (degradation), θεωρώντας πως **τα τμήματα ηχογραφημένης φωνής χάνουν την φυσικότητά τους, ανάλογα με τον βαθμό ψηφιακής επεξεργασίας τους.**

Στην συνέχεια, με την βοήθεια wide band σπεκτρογραμμάτων, έγινε η εξαγωγή των διφώνων από τις ηχογραφημένες λέξεις.

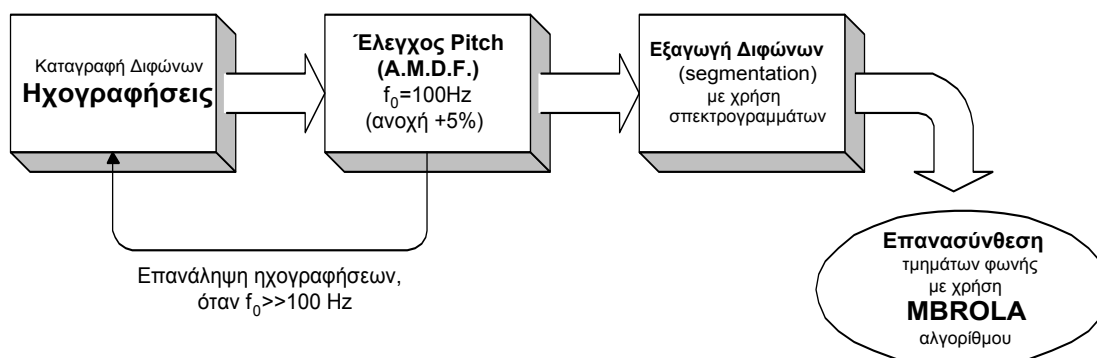
Κάθε δίφωνο, περιείχε steady-state τμήμα του αρχικού φθόγγου, μετάβαση στον επόμενο και steady-state τμήμα του επόμενου φθόγγου. Σε ένα αρχείο κειμένου, σημειωνόταν το δείγμα από το οποίο ξεκινούσε ο 1ος φθόγγος του διφώνου, το δείγμα που τελείωνε ο 2ος φθόγγος του διφώνου και το δείγμα που οριακά χώριζε τους δύο φθόγγους του διφώνου

⁴ Η AMDF βασίζεται στο γεγονός ότι εάν το σήμα $x(n)$ είναι περιοδικό με περίοδο P , τότε η $d(n)=x(n)-x(n-k)$ μηδενίζεται για $k=0, \pm P, \pm 2P, \dots$. Συνεπώς, το μέσο πλάτος της $d(n)$, σε ένα πεπερασμένο χρονικό διάστημα, θα ελαχιστοποιείται όταν το k παίρνει τιμές κοντά στην περίοδο P . Αποδεικνύεται πως η εύρεση της περιόδου με την τεχνική αυτή, καταλήγει στην εύρεση των ελαχίστων της συνάρτησης, $\gamma_n(k)=[R_n(0)-R_n(k)]^{1/2}$, όπου $R_n(k)$ η αυτοσυσχέτιση μικρού χρόνου.

(αρχή, τέλος, μέση διφώνου). Το ποσοστό steady-state τμήματος από κάθε φθόγγο που εξαγόταν σε κάθε δίφωνο, καθοριζόταν από την γειτονιά του κάθε φθόγγου. Σε C-V (Consonant-Vowel) ή V-C (Vowel-Consonant) περιοχές, η συνάρθρωση είναι εντονότερη στο μέσον των φωνηέντων απ' ό,τι στο τέλος των φωνηέντων (για C-V) ή απ' ό,τι στην αρχή (για V-C), όπως φαίνεται και στο σχήμα που ακολουθεί. Συνεπώς, σ' αυτές τις γειτονιές, το σημείο διαχωρισμού (cuttpoint), δεν θα πρέπει να είναι το μέσον των φωνηέντων. Σε μικτές όμως περιοχές, (C-V-C), το επιλεγόμενο, σημείο διαχωρισμού ήταν το μέσον των φωνηέντων.

Σε συνδυασμούς ευρέως μεταβαλλόμενων (highly transient) φθόγγων, όπως στους κροτικούς (plosives), ο διαχωρισμός των διφώνων που περιέχουν τους φθόγγους αυτούς, είναι προβληματικός. Στην εργασία αυτή, οι στιγμιαίοι (κροτικοί) φθόγγοι, αντιμετωπίστηκαν ως άθροισμα ενός διαστήματος παύσης και ενός διαστήματος 'έκρηξης'. Στα δίφωνα της μορφής /x-p/, όπου x είναι οποιοσδήποτε φθόγγος και p οποιοδήποτε στιγμιαίο, περιλαμβάνεται το διάστημα σιωπής, ενώ στα δίφωνα της μορφής /p-x/ περιλαμβάνεται το διάστημα έκρηξης.

Το διάγραμμα που παρουσιάζει τις διεργασίες για την εξαγωγή (segmentation) των διφώνων, παρουσιάζεται παρακάτω:



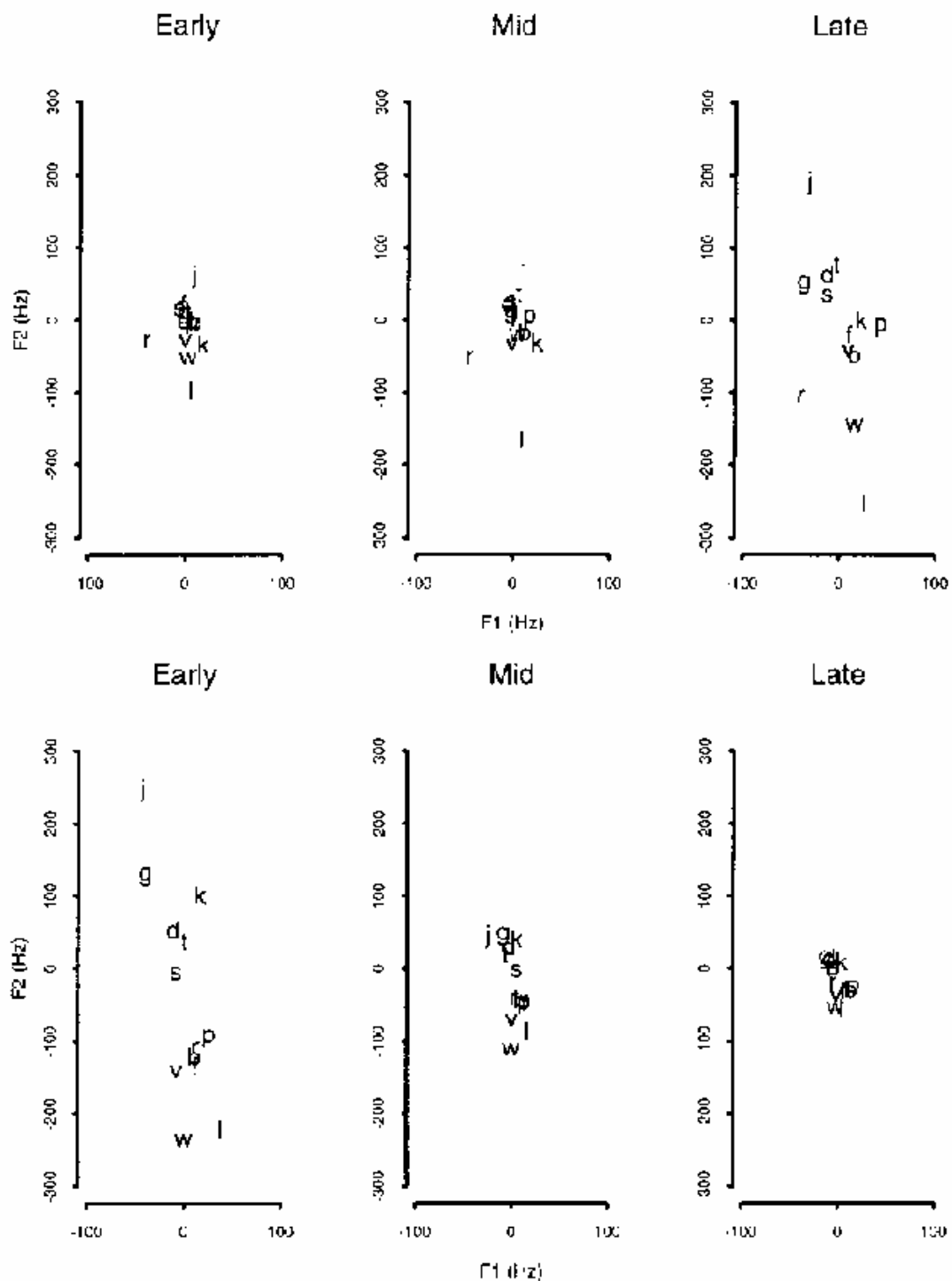
[περιγραφή της διαδικασίας παραγωγής των διφώνων, τα οποία χρησιμοποιούνται στην τεχνική αλυσιδωτής σύνθεσης (concatenation synthesis), όπως υλοποιείται στον Αίσωπο]

Η συνολική πληροφορία που αναγκαστήκαμε να επεξεργαστούμε, έφτασε το 1 Gbyte, για να εξαχθούν δίφωνα, συνολικής χωρητικότητας 1.5 Mbyte ! Η διαδικασία επεξεργασίας (έλεγχος pitch και εξαγωγή διφώνων), ήταν ιδιαίτερος επίπονη και χρονοβόρα, καθώς κράτησε δύο μήνες.

Η τελική βάση διφώνων, ανασυνθέθηκε με οδηγό τον αλγόριθμο MBROLA, όπως περιγράφηκε σε προηγούμενο κεφάλαιο, σε συνεργασία με το Πανεπιστήμιο της Mons, (καθηγητής T.Dutoit, Βέλγιο) και διανέμεται έκτοτε, στο διαδίκτυο από το site του εργαστηρίου Μαγνητικού Συντονισμού της σχολής μας⁵ αλλά και από το αντίστοιχο site του Πανεπιστημίου της Mons⁶.

⁵ <http://mri.ee.auth.gr/synthesis>

⁶ <http://tcts.fpms.ac.be/mbrola.htm>



[στο σχήμα φαίνεται η επίδραση στις συχνότητες διαμόρφωσης F1,F2, των φωνηέντων της Αγγλικής από επιλεγμένα σύμφωνα, όταν προηγούνται των φωνηέντων (κάτω σχήμα) ή όταν ακολουθούν τα φωνήεντα (επάνω σχήμα). Φαίνεται καθαρά, πως το μέσον (middle) των φωνηέντων είναι σημείο μεγαλύτερης συνάρθρωσης απ'ότι η αρχή (Early, επάνω σχήμα) σε VC 'γειτονιές' ή απ'ότι το τέλος (Late, κάτω σχήμα) σε CV 'γειτονιές', (The Bell Labs Approach, 1998)]

Κεφάλαιο τέσσερα

Φωνηματική Μεταγραφή

Στην ενότητα αυτή, θα συνοψίσουμε το σύνολο των γραμματικών κανόνων, όπως εφαρμόστηκαν στον Αίσωπο, για την μετατροπή των γραμμάτων των λέξεων, στα φωνηματικά σύμβολα του αλφαβήτου S.A.M.P.A. Οι κανόνες μεταγραφής είναι απαραίτητοι, καθώς δεν υπάρχει αμοιμονοσήμαντη σχέση μεταξύ γραμμάτων και φωνημάτων. Το γεγονός αυτό, οφείλεται στην πιο γοργή ιστορική εξέλιξη του Προφορικού Λόγου σε σχέση με τον Γραπτό.

Συνεπώς, η φωνηματική μεταγραφή, είναι ένα σύνθετο πρόβλημα.

4.1 Τεχνικές Φωνηματικής Μεταγραφής - Επιλογή μεθόδου για τον “Αίσωπο”

Ωστόσο, μπορούμε να πούμε, πως το πρόβλημα της μετατροπής των γραμμάτων στα αντίστοιχα φωνήματα, για την Ελληνική Γλώσσα επιλύεται λιγότερο δύσκολα απ’ό,τι για τις υπόλοιπες Ευρωπαϊκές Γλώσσες. Στα Ελληνικά, γράφουμε τις λέξεις, σχεδόν όπως τις προφέρουμε, με λίγες εξαιρέσεις. Επιπλέον, οι ομοιομορφικές (homomorphic)¹ λέξεις της Ελληνικής είναι ελάχιστες, απλοποιώντας έτσι το πρόβλημα.

Στα πλαίσια της κατασκευής του “Αίσωπου” για την Ελληνική Γλώσσα, προτιμήσαμε τον σχηματισμό ενός συνόλου κανόνων σε μορφή δένδρου, σε συνδυασμό με ένα σύνολο λεξικών, που περιέχουν μορφήματα (morphemes) των λέξεων, που αποτελούν εξαιρέσεις στους παραπάνω κανόνες. Επομένως, η προσέγγιση στον Αίσωπο, του προβλήματος της φωνηματικής μεταγραφής βασίζεται περισσότερο σε κανόνες και λιγότερο στις εξαιρέσεις τους (**rule-based transcription**). Η μεθοδολογία αυτή, έχει εφαρμοστεί σε πληθώρα συστημάτων με επιτυχία (Hunnicut, 1980).

Μια άλλη τεχνική, θα ήταν η αποθήκευση των πιο συχνά χρησιμοποιούμενων λέξεων της Γλώσσας, υπό μορφή μορφημάτων σε λεξικό, μαζί με την φωνηματική μεταγραφή τους (**dictionary-based transcription**). Η μεθοδολογία αυτή, εφαρμόστηκε στο TTS σύστημα που κατασκευάστηκε στο M.I.T. (MITALK, Allen (1987)), όπου σχηματίστηκε ένα λεξικό από 12.000 μορφήματα λέξεων, μαζί με την φωνηματική τους μεταγραφή. Το TTS σύστημα επίσης, της AT&T, χρησιμοποιεί λεξικό 43.000 φωνημάτων.

Στις dictionary-based υλοποιήσεις, χρησιμοποιούνται και πάλι κανόνες, στην περίπτωση που κάποια λέξη, δεν έχει συμπεριληφθεί έστω και ως μόρφημα στο λεξικό. Συνεπώς, οι

¹ οι λέξεις, που ενώ γράφονται το ίδιο προφέρονται διαφορετικά π.χ. η λέξη record ως ουσιαστικό τονίζεται στην πρώτη συλλαβή, ενώ ως ρήμα, τονίζεται στην τελευταία

τεχνικές αυτές, είναι μεν αποτελεσματικές, αλλά ταυτόχρονα απαιτούν πολύ χρόνο έτσι ώστε να σχηματιστούν τα λεξικά φωνηματικής μεταγραφής των μορφημάτων.

Εκτός από τις παραπάνω ντετερμινιστικές τεχνικές, έχουν αναφερθεί στην βιβλιογραφία τεχνικές φωνηματικής μεταγραφής, βασισμένες στην εκπαίδευση ενός συστήματος, με χρήση Hidden Markov Models (Van Coile, 1991) ή με χρήση νευρωνικών δικτύων (Ainsworth & Pell, 1989). Το πλεονέκτημα αυτών των τεχνικών, είναι ότι δίνουν λύσεις, χωρίς την χρήση εξειδικευμένων γραμματολογικών γνώσεων και κανόνων.

Ωστόσο, στον “Αίσωπο”, προτιμήθηκε ένα σύνολο προσεκτικά σχεδιασμένων κανόνων μαζί με τα λεξικά των εξαιρέσεων, καθώς αυτή η μεθοδολογία έδωσε σχετικά γρήγορα, αξιόπιστα αποτελέσματα, αποδεικνύοντας έτσι, πως ταιριάζει περισσότερο στην δομή και στις ιδιαιτερότητες της Ελληνικής Γλώσσας.

4.2 Δυνατότητες του Αλγορίθμου Φωνηματικής Μεταγραφής

Ο αλγόριθμος που εφαρμόστηκε και που θα περιγραφεί στην επόμενη ενότητα, έχει τις παρακάτω δυνατότητες:

- i. μετατροπή των συνδυασμών αυ,ευ σε αν,εν ή af,ef ανάλογα με τον φθόγγο που ακολουθεί π.χ. τον συνδυασμό <αυ> στην λέξη ‘αυγή’, ο αλγόριθμος τον μετατρέπει σε /av/, ενώ τον συνδυασμό <αυ> στην λέξη ‘αυθεντικός’ τον μετατρέπει σε /af/
- ii. απαλοιφή γραμμάτων όπου χρειάζεται π.χ. στην λέξη εύφορος το <υ> δεν προφέρεται
- iii. μετατροπή του <σ> σε /z/ όπου χρειάζεται π.χ. στην λέξη ‘σβήνω’ το <σ> προφέρεται /z/ (/zvino/).
- iv. μετατροπή του συνδυασμού <μπ> σε /mp/ όπως στην λέξη ‘τέμπο’ (→/tempo/) ή σε /b/ όπως στη λέξη ‘μπαίνω’ (→/beno/).
- v. μετατροπή του <γγ> σε /g/, /ng/, /nG/, /Gn/ όπως στις λέξεις ‘γκολ’ (/gol/), ‘φεγγάρι’ (/fengari/), ‘συγγραφέας’ (/sinGrafeas/), ‘συγγνώμη’ (/siGnomi/) αντίστοιχα.
- vi. μετατροπή του <ντ> σε /d/, /nd/, /nt/ όπως στις λέξεις ‘ντύνω’ (/dino/), ‘αντάρτης’ (/andartis/) και ‘κάντο’ (/kanto/) αντίστοιχα.
- vii. μετατροπή του <γχ> σε /nx/ ή /nC/ όπως στις λέξεις ‘συγχέω’ (/sinCeo/) ή ‘συγχωρώ’ (/sinxoro/).
- viii. μετατροπή του <γκτ> σε /nkt/ και του <γξ> σε /nks/ όπως στις λέξεις ‘άτεγκτος’ (/atenktos/) και ‘σύσφιγξη’ (/sisfinksi/) αντίστοιχα.

ix. επιτυχής χειρισμός του φαινομένου της **συνίζησης**, κατά το οποίο οι δίφθογγοι με πρώτο προταχτικό το /i/, λειτουργούν ως καταχρηστικοί δίφθογγοι με συνέπεια το /i/ να

μετατρέπεται στο μισόφωνο /j/. Π.χ. ο δίφθογγος <ιο> στη λέξη ‘αναλόγιο’ προφέρεται /io/ (/analojjio/), ενώ στην λέξη ‘μουράγιο’ προφέρεται /jo/ (/murajjo/). Η στην λέξη ‘πιστικός’, ο συνδυασμός <ιε> προφέρεται /ie/ (piestikos/), ενώ στην λέξη ‘πιες’, προφέρεται /je/ (/pjes/).

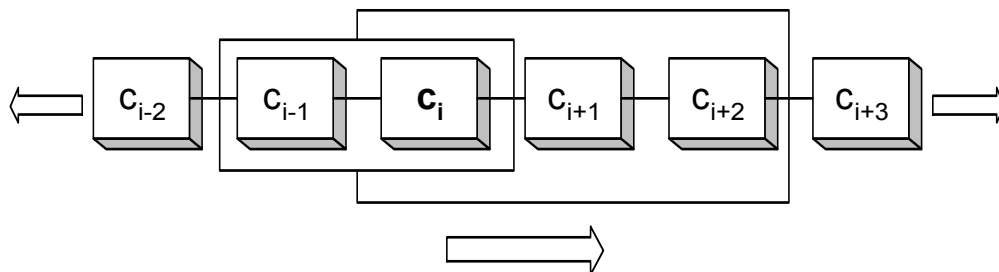
Ο αλγόριθμος φωνηματικής μεταγραφής, οφείλει επίσης να μετατρέπει τα νούμερα σε λέξεις, στο σωστό γένος και στην σωστή πτώση, ενώ πρέπει να διαθέτει ικανότητες μετατροπής των συντομογραφιών σε ολόκληρες λέξεις.

Αν και η μετατροπή αριθμών σε λέξεις, δεν είναι δύσκολη υπόθεση, η σωστή κλίση τους, ανάλογα με τις απαιτήσεις του κειμένου, αποτελεί ξεχωριστό πρόβλημα².

4.3 Περιγραφή Αλγορίθμων Φωνηματικής Μεταγραφής

Στο αρχικό στάδιο της επεξεργασίας, κάθε αλγόριθμος φωνηματικής μεταγραφής μετατρέπει τα νούμερα και τις συντομογραφίες σε λέξεις. Ο Αίσωπος, δεν ενσωματώνει ακόμη τέτοια χαρακτηριστικά. Στο στάδιο της προεπεξεργασίας, ο Αίσωπος μετατρέπει όλους τους χαρακτήρες σε πεζούς, έτσι ώστε στην συνέχεια, να απλοποιηθεί η φωνηματική μεταγραφή.

Στην συνέχεια, εξετάζεται κάθε χαρακτήρας, σε σχέση με τους γειτονικούς του, σύμφωνα με το σχήμα:



[εξέταση κάθε χαρακτήρα c_i , σε σχέση με τους γειτονικούς του]

Ο χαρακτήρας c_{i-1} , αντιπροσωπεύει είτε τον αμέσως προηγούμενο χαρακτήρα της φράσης, είτε τον τελευταίο χαρακτήρα της αμέσως προηγούμενης λέξης, όταν ο προηγούμενος χαρακτήρας του c_i είναι το κενό.

Ο χαρακτήρας c_i , εξετάζεται με βάση την σειριακή εφαρμογή των κανόνων που ακολουθούν, σε σχέση με τους γειτονικούς του χαρακτήρες. Όταν κάποιος από τους παρακάτω κανόνες εφαρμόζεται, ο έλεγχος του προγράμματος περνά στον επόμενο ή στον μεθεπόμενο χαρακτήρα.

² Ενώ ασχοληθήκαμε με την μετατροπή αριθμών σε λέξεις, με δεδομένα το γένος και την πτώση, μέχρι την στιγμή συγγραφής της εργασίας αυτής, δεν είχαμε ολοκληρώσει τους αλγορίθμους πρόβλεψης του γένους και της πτώσης ενός αριθμού, μέσα από το κείμενο.

Αξίζει να τονίσουμε ότι μέρος των παραπάνω κανόνων, με τον ίδιο φορμαλισμό (δένδρο), εφαρμόζεται στην συμβολοσειρά των φωνημάτων που προκύπτει, μετά την φωνηματική μεταγραφή, διορθώνοντας λάθη του αλγορίθμου φωνηματικής μεταγραφής.

Η οργάνωση των κανόνων, με την παρακάτω μορφή δένδρου, εξυπηρετεί την εύκολη συμπλήρωσή τους και βοηθά στην γενικότερη εποπτεία του προγράμματος. Ο μικρός αριθμός φωνημάτων του SAMPA αλφάβητου, σε συνδυασμό με το γεγονός ότι οι δομικές μονάδες του συνθέτη φωνής του Αισωπου είναι τα δίφωνα (=συνδυασμοί φωνημάτων), διευκολύνει την γρήγορη μεταγραφή ολόκληρης της φράσης. Εάν είχαμε μεγαλύτερες δομικές μονάδες, π.χ. τρίφωνα, θα χρειαζόμασταν μεγαλύτερο πλήθος κανόνων και επομένως η διαδικασία μεταγραφής θα ήταν πιο αργή.

Κατά την διάρκεια της εξέτασης του χαρακτήρα c_i με βάση το σύνολο των παρακάτω κανόνων, γίνεται, όπου απαιτείται, παραπομπή του προγράμματος σε λεξικά, που περιέχουν μορφήματα (morphemes) των λέξεων που αποτελούν εξαιρέσεις στον κανόνα που εκείνη την στιγμή εξετάζεται. Εάν βρεθεί μόρφημα του λεξικού εξαίρεσης, στο σημείο του κειμένου εισόδου που βρίσκεται ο έλεγχος του προγράμματος, ενεργοποιείται η φωνηματική μεταγραφή που υπαγορεύει το λεξικό εξαίρεσης για το συγκεκριμένο μόρφημα.

Τα λεξικά εξαίρεσης, είναι οργανωμένα ανάλογα με το φαινόμενο το οποίο αντιμετωπίζουν (π.χ. άλλο λεξικό για το φαινόμενο της συνίζησης και άλλο λεξικό για τον χειρισμό του <γγ> σε /g/ ή /G/). Η προσεκτική κατηγοριοποίηση των λεξικών, ωφελεί στην πιο οργανωμένη ενημέρωσή τους, ενώ παράλληλα ελαχιστοποιεί τον χρόνο επεξεργασίας, αφού κάθε φορά εξετάζεται το κατάλληλο λεξικό και όχι το σύνολο των εξαιρέσεων.

Τα λεξικά μορφημάτων εξαιρέσεων που χρησιμοποιούνται στον “Αίσωπο”, χωρίζονται σε κατηγορίες, περιλαμβάνοντας τις περιπτώσεις που:

1. το <μπ> μετατρέπεται σε /mp/ (και όχι σε /b/)
2. το <ντ> μετατρέπεται σε /nt/ (και όχι σε /d/)
3. το <γγ> μετατρέπεται σε /G/ (και όχι σε /g/)
4. ο δίφθογγος που αντιστοιχεί σε /ia/ μετατρέπεται σε /ia/
5. ο δίφθογγος που αντιστοιχεί σε /io/ μετατρέπεται σε /io/
6. ο δίφθογγος που αντιστοιχεί σε /iu/ μετατρέπεται σε /iu/
7. ο δίφθογγος που αντιστοιχεί σε /ie/ μετατρέπεται σε /ie/
8. ο δίφθογγος που αντιστοιχεί σε /ia/ μετατρέπεται σε /ja/
9. ο δίφθογγος που αντιστοιχεί σε /io/ μετατρέπεται σε /jo/

Συνεπώς, χρησιμοποιούνται 9 λεξικά-υπορουτίνες στο στάδιο φωνηματικής μεταγραφής. Τα λεξικά 4-7 χρησιμοποιούνται όταν η προκαθορισμένη (default) τιμή μετατροπής για το <ι> (οι,ει κλπ) είναι το /j/, ενώ τα λεξικά 8-9 όταν η default τιμή μετατροπής για το <ι> (οι,ει, κλπ) είναι το /i/ , όπως π.χ. στο τέλος των λέξεων που τελειώνουν σε -ριο (π.χ. ‘παρατηρητήριο’).

Τα λεξικά 4-9 χρησιμοποιούνται στον χειρισμό του φαινομένου της συνίζησης, ο οποίος γίνεται στο τέλος του δένδρου κανόνων που ακολουθεί.

Η στοιχειώδης ενημέρωση των λεξικών εξαιρέσεων, με την βοήθεια ηλεκτρονικού λεξικού, οδήγησε σε ποσοστά επιτυχούς φωνηματικής μεταγραφής, της τάξης του 92 % !

Είναι βέβαιο, πως η περαιτέρω ενημέρωση των λεξικών θα βελτιώσει την επίδοση του “Αισωπου”.

Ωστόσο, ο παραπάνω αλγόριθμος, δεν μπορεί να κάνει τίποτε στις ελάχιστες ομοιομορφικές λέξεις της Ελληνικής, όπως στην λέξη ‘άδεια’ η οποία μπορεί να προφέρεται /aDia/ όπως π.χ. στην πρόταση << πήρα πενθήμερη ‘άδεια’>>, ή /aDja/ όπως στην πρόταση <<έχω την τσέπη άδεια από λεφτά>>.

Οι ομοιομορφικές λέξεις, απαιτούν υψηλότερα επίπεδα σημασιολογικής επεξεργασίας (semantics).

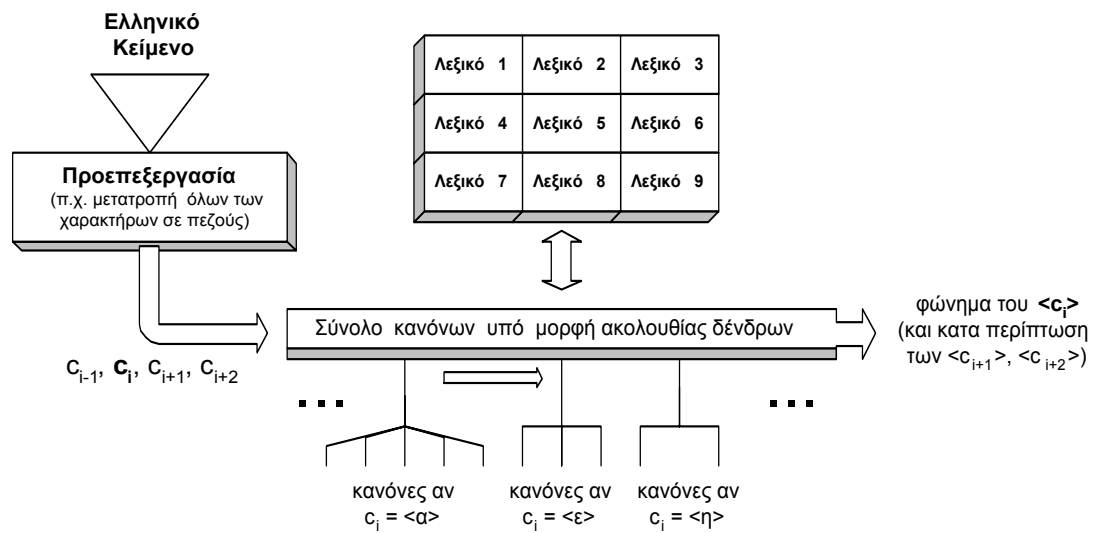
C _{i-1}	C _i	C _{i+1}	C _{i+2}	Φωνηματική Μεταγραφή
	τ	σ,ς		ts
	τ	ζ		tz
	τ			t
	α	ϊ,ί,ή,η		ai
	α	ι,ί		e
	α	υ,ύ	ηχηρό σύμφωνο ή φωνήεν	av
	α	υ,ύ	άηχο σύμφωνο	af
	α			a
	ά	ϊ,ί,ή,η,ι		ai
	ά	υ,ύ	ηχηρό σύμφωνο ή φωνήεν	av
	ά	υ,ύ	άηχο σύμφωνο	af
	ά			a

	ε	ĩ,ĩ,ή,η		ei
	ε	ί,ι*		i
	ε	υ,ύ	ηχηρό σύμφωνο ή φωνήεν	ev
	ε	υ,ύ	άηχο σύμφωνο	ef
	ε			e
	έ	ĩ,ĩ,ή,η,ι		ei
	έ	υ,ύ	ηχηρό σύμφωνο ή φωνήεν	ev
	έ	υ,ύ	άηχο σύμφωνο	ef
	έ			e
	η	υ,ύ	ηχηρό σύμφωνο ή φωνήεν	iv
	η	υ,ύ	άηχο σύμφωνο	if
	η			i
	ή	υ	ηχηρό σύμφωνο ή φωνήεν	iv
	ή	υ	άηχο σύμφωνο	if
	ή			i
	ω			o
	ώ			o
	ο	υ,ύ		u
	ο	ĩ,ĩ,ũ,ũ		oi
	ο	ί,ι		i
	ο			o
	ό			o
	υ	ι		i
	υ			i
	ύ	ι		i
	ύ			i
	ϣ	ι		i
	ϣ			i
	β			v
	φ			f
	δ			D
	ζ			z

	θ			T
	λ			l
	π			p
	ρ			r
	ξ			ks
	ψ			ps
	σ,ς	φωνήεν, άηχο σύμφωνο, λ,σημείο στίξης		s
	σ,ς	ηχηρό σύμφωνο		z
	κ	α,ά,ο,ό		k
	κ	ε,η,ή,ι,ί,υ,ύ		c
	κ			k
	χ	α,ά,ο,ό		x
	χ	ε,η,ή,ι,ί,υ,ύ		C
	χ			x
	μ	π [*]	(τ)	mp
	μ	π		b
	μ			m
	ν	κ		ng
	ν	π		mp
αρχή φράσης	ν	τ	≠ζ, ≠σ	d
ρ	ν	τ [*]	≠ζ, ≠σ	d
όχι αρχή	ν	τ [*]	≠ζ, ≠σ	nd
	ν	τ [*]	≠ζ, ≠σ	nt
	ν			n
	γ	χ, ξ, κ (c _{i+2} =τ)	τ (c _{i+1} =κ)	n
	γ	έ,ε,ή,η,ί,ι,υ,ύ		jj
αρχή φράσης	γ	κ		g
όχι αρχή	γ	κ		ng
	γ	γ	v	Gn
	γ	γ [*]		ng
	γ	γ [*]		nG

	γ	γ	ρ		gr
	γ				G
	ί				i
	ι	σύμφωνο			i
αρχή λέξης	ι				i
όλα τα σύμφωνα	ι [*]	φωνήεν			j
όλα τα σύμφωνα	ι [*]	φωνήεν			i (ανάλογα με τις εξαιρέσεις των λεξικών)
	ι				j (default τιμή)
	“				null
	space				null
	-				—
	‘				null
	·				— —
	;				— —
	?				— —
	:				— —
	!				— —
	,				—
	υπόλοι- ποι χαρα- κτήρες				?

[οι κανόνες εφαρμόζονται σειριακά, με ‘συμβουλή’ των λεξικών εξαιρέσεων, όπου χρειάζεται. Η παραπομπή σε κάποιο από τα 9 λεξικά εξαιρέσεων, συμβολίζεται με αστερίσκο στον παραπάνω πίνακα. Όταν κάποιος από τους παραπάνω κανόνες εφαρμόζεται, με συνέπεια την φωνηματική μεταγραφή του c_i (και των c_{i+1}, c_{i+2} αν χρειαστεί) ο έλεγχος του προγράμματος, περνά στους επόμενους χαρακτήρες και η εξέταση των κανόνων ξεκινά ξανά από την αρχή]



[σχηματική αναπαράσταση του αλγορίθμου φωνηματικής μεταγραφής, που ακολουθείται στον “Αίσωπο”]

Κεφάλαιο 5

Μοντέλα Προσωδίας

Στην ενότητα αυτή, θα ξεκαθαρίσουμε την έννοια της προσωδίας και θα υπογραμμίσουμε τις παραμέτρους που την καθορίζουν. Στην συνέχεια, θα περιγράψουμε τα προσωδικά μοντέλα που χρησιμοποιούνται στο Αίσωπο.

5.1 Βασικές έννοιες

Ο τονισμός των λέξεων καθώς και τα σημεία στίξης, αποτελούν την προσωδία μιας πρότασης, και ανάγονται στα παραγλωσσικά σημεία που καθορίζουν ή/και τροποποιούν την σημασία των γλωσσικών στοιχείων, επιτρέποντας την έναρξη, διατήρηση ή διαφοροποίηση της προφορικής, κυρίως, (αλλά και της γραπτής) επικοινωνίας.

Πρακτικά, η προσωδία καθορίζει την ένταση, την διάρκεια και την βασική συχνότητα κάθε φωνήματος, μέσα σε μια πρόταση, με υπερτμηματικό τρόπο, αφού αντικείμενο μελέτης της είναι τα “προσωδήματα”, εκφραστικές ενότητες που ως προς το μήκος τους ξεπερνούν τα φωνήματα, ενώ οι κατηγορίες που αυτά εντάσσονται είναι πολλές και διαχωρίζονται με βάση τις παύσεις, τον ρυθμό, το είδος τονισμού κλπ. Η μελέτη τους, μπορεί να γίνει σε συντακτικό επίπεδο (αφορά τον επιτονισμό, τον τονισμό δηλαδή με μεταβολές του ύψους-pitch), σε μορφοσυντακτικό επίπεδο (ο τόνος της κάθε λέξης) ή σε μορφολογικό επίπεδο (π.χ. εάν ο τόνος βρίσκεται στην 1^η, 2^η ή 3^η συλλαβή). Η προσωδία είναι συνιστώσα της Φωνολογίας, και η επιτυχής πρόβλεψη των προσωδικών παραμέτρων (ένταση, διάρκεια, βασική συχνότητα) κάθε φωνήματος, ενός Γραπτού κειμένου, θα προσδώσει “φυσικότητα” στην εκφώνηση του κειμένου, ενώ ταυτόχρονα θα συμβάλει στην κατανόηση του νοήματος-μηνύματος της προφερόμενης πρότασης.

5.1.1 ΤΟΝΙΣΜΟΣ - Τόνος Δυναμικός, Επιτόνηση, Ποιότητα-Διάρκεια

Τονισμός είναι η χρήση διακριτικών σημείων-σημαδιών, που τοποθετούνται στα γράμματα που αντιπροσωπεύουν κάποια φωνηεντικά φωνήματα (φωνήεντα). Στην ΝεοΕλληνική, μετά την μεταρρύθμιση του 1976, οπότε καθιερώθηκε ως επίσημη γλώσσα του Ελληνικού Κράτους η δημοτική, χρησιμοποιείται ως τόνος των λέξεων, μόνο η οξεία. Για ιστορικούς λόγους αναφέρουμε ότι έως τότε, από τα αρχαία χρόνια και σε όλες τις φάσεις της ιστορίας της γλώσσας (διάλεκτοι των αρχαίων χρόνων όπως η αττική, η δωρική κ.α., στη γλώσσα της ελληνιστικής περιόδου, των ρωμαϊκών και κατόπιν των βυζαντινών χρόνων, στη γλώσσα των ‘λογίων’ των μεταβυζαντινών χρόνων, στην

ελληνική επί της Τουρκοκρατίας και κατόπιν σ'εκείνην της Ελληνικής Επανάστασεως, έως και στην επίσημη χρήση της 'καθαρεύουσας' μέχρι το 1976), οι τόνοι της Ελληνικής είναι τρεις: η οξεία, η βαρεία και η περισπωμένη. Η θέση τους ήταν σε κάποια από τις 3 τελευταίες συλλαβές των λέξεων, δηλ. την λήγουσα, την παραλήγουσα και την προπαραλήγουσα.

Όσο αφορά τη χρήση της περισπωμένης και της οξείας (ή βαρείας, με χρήση σχεδόν ταυτόσημη στα νεότερα χρόνια εκτός από εκείνην των αρχαίων χρόνων όπου υπήρχαν κάποιες διαφοροποιήσεις) θεσπίστηκαν κανόνες, που αφορούσαν την παρουσία στις συλλαβές μακρών και βραχέων φωνηέντων. Έτσι, αξιομνημόνευτοι είναι οι κανόνες όπως 'μακρό προ μακρού τονίζεται', 'λήγουσα βραχεία, παραλήγουσα μακρά, τονιζόμενη δέχεται περισπωμένη', 'βραχεία πάντοτε τονίζονται' κ.α.

Φυσικά, αυτές οι διακρίσεις σε φωνήεντα μακρά και βραχεία ήταν λειτουργικές στα αρχαία χρόνια, ενώ πλέον δεν υφίσταται.

- **Τόνος Δυναμικός, Επιτόνιση, Ποιότητα-Διάρκεια**

Από αυστηρά γλωσσολογική άποψη, η προσωδία έχει 3 χαρακτηριστικά στοιχεία, που είναι ο **δυναμικός τόνος** (ή **τόνος ενέργειας**, σχετικά με τη μεγαλύτερη ή μικρότερη ένταση που έχει ο αέρας της εκπνοής), ο **τόνος της επιτόνισης** (ή τόνος του ύψους, σχετικά με τη μεγαλύτερη ή μικρότερη βασική συχνότητα -pitch) και η **διάρκεια ή ποιότητα** που σχετίζεται με το μεγαλύτερο ή μικρότερο "κράτημα" του ήχου κάποιου φωνήματος.

Ειδικά, η επιτόνιση παρουσιάζει ενδιαφέρον μελέτης, όχι μόνο σε επίπεδο λέξης, αλλά και σε επίπεδο φράσης (**intonation**), λόγω της επίδρασής της στο νόημα της φράσης. Έτσι, το ανέβασμα του ύψους της φωνής (pitch) στο τέλος της πρότασης δείχνει ερώτηση, το κατέβασμά της δείχνει κατάφαση ή άρνηση, όπως επίσης το σταμάτημα σε κάποιο σημείο και η συνέχιση προς μεγαλύτερες ή μικρότερες τιμές pitch, δείχνουν τη διαφοροποίηση της κύριας από τις δευτερεύουσες προτάσεις, ανάλογα, φυσικά και με το περιεχόμενο του λόγου που περιέχεται σ'αυτές.

5.1.2 Τα Σημεία στίξης

Πρόκειται για τα σημάδια εκείνα, που τοποθετούμενα στο γραπτό λόγο, δείχνουν τη διαφοροποίηση των κυρίων από τις δευτερεύουσες προτάσεις, καθώς και τη γενικότερη ή ειδικότερη σχέση που έχουν τα γλωσσικά στοιχεία κατά τους συνδυασμούς τους.

Ωστόσο, θεμελιώδης είναι η σημασία τους και στον προφορικό λόγο:

- Η τελεία < . > : χρησιμοποιείται στο τέλος των προτάσεων που έχουν πλήρες νόημα, ενώ κατά την εκφορά του λόγου, η φωνή σταματάει και ο τόνος της (pitch) “πέφτει”.
- Η άνω τελεία < ' > : μ’αυτήν δείχνουμε σταμάτημα στη φωνή, μικρότερο της τελείας και μεγαλύτερο του κόμματος.
- Το κόμμα < , > : χρησιμοποιείται πιο συχνά σε σχέση με τα υπόλοιπα σημεία στίξης, για να δειχθεί ένα μικρό σταμάτημα της φωνής. Με αυτό χωρίζονται λέξεις ασύνδετες που είναι όμως ίδια μέρη του λόγου, οι όμοιες ασύνδετες προτάσεις και οι δευτερεύουσες από τις κύριες προτάσεις.
- Το ερωτηματικό < ; > : η παρουσία του οδηγεί στο ‘ανέβασμα’ της φωνής στο τέλος της ερωτηματικής -επιβεβαιωτικής πρότασης (π.χ. “θα πας στο χορό το βράδυ;”) ή στο ‘ανέβασμα’ (=μεγαλύτερο pitch) της φωνής και στην αρχή και στο τέλος ερωτηματικής μη επιβεβαιωτικής πρότασης (π.χ. “Πού θα πας το βράδυ;”).
- Το θαυμαστικό < ! > : στον προφορικό λόγο, εκφράζεται με τον ‘τμηματικό’ τονισμό των λέξεων που προηγούνται της παρουσίας του στο τέλος των προτάσεων.
- Η διπλή άνω και κάτω τελεία < : > : λειτουργεί στον προφορικό λόγο ως παύση, διάρκειας μεταξύ της παύσης του κόμματος και της παύσης της τελείας.
- Η παρένθεση < () > : στον προφορικό λόγο, μπορεί να λειτουργήσει επεξηγηματικά, απαιτώντας μεταβολή της μέσης βασικής συχνότητας.
- Τα αποσιωπητικά < ... > : στον προφορικό λόγο δείχνουν συγκίνηση, θαυμασμό, ή κάποιο άλλο συναίσθημα.
- Η παύλα, η διπλή παύλα, και τα εισαγωγικά < - -- << >> “ ” ‘ ’ > : στον προφορικό λόγο λειτουργούν, μόνο όταν η φράση που περικλείουν έχει επεξηγηματική αξία.

5.2 Τα Προσωδικά Μοντέλα στον “Αίσωπο”

Από τα παραπάνω γίνεται κατανοητό ότι η Προσωδία ενός κειμένου, σχετίζεται με την ένταση κάθε φθόγγου (δυναμικός τόνος ή τόνος ενέργειας), την βασική του συχνότητα (επιτόνιση) και την διάρκειά του (ποιότητα-διάρκεια). Επομένως, το σύστημα Αυτόματης Ανάγνωσης και Εκφώνησης, εκτός από την μεταγραφή των γραμμάτων στα αντίστοιχα

φωνήματα, οφείλει να προβλέπει τις διάρκειές τους, την βασική τους συχνότητα (pitch) και πώς αυτή μεταβάλλεται στα πλαίσια της φράσης (intonation), καθώς επίσης και να υπολογίζει την ένταση του κάθε φθόγγου.

Η ένταση του κάθε φθόγγου, εξαρτάται από το πλάτος της κυματομορφής που τον αντιπροσωπεύει, σε μια μέθοδο σύνθεσης φωνής, με χρήση προηχογραφημένων μονάδων (concatenation). Στις περισσότερες concatenation υλοποιήσεις, η παράμετρος της έντασης, λαμβάνεται έμμεσα υπόψιν, καθώς όπου απαιτείται αύξηση της έντασης του φθόγγου, αυξάνεται ταυτόχρονα η διάρκεια και το pitch του, χωρίς να επηρεάζεται το πλάτος της κυματομορφής στην οποία ο φθόγγος αντιστοιχεί. Αυτή η λογική ακολουθήθηκε και στον “Αίσωπο”, καθώς απλοποιεί την απαραίτητη Επεξεργασία Φυσικής Γλώσσας και κυρίως την διαδικασία Σύνθεσης, περιορίζοντάς την στον χειρισμό δύο (διάρκεια, pitch) από τις τρεις προσωδικές παραμέτρους (ένταση, διάρκεια, pitch).

Είναι προφανές, πως η διαδικασία εκφώνησης ενός κειμένου, είναι μια περίπλοκη διαδικασία, καθώς έχει να κάνει με ένα μεγάλο σύνολο κανόνων επικοινωνίας που ονομάζεται Γλώσσα, ενώ ταυτόχρονα, απαιτεί εμπειρία (γι’αυτό άλλωστε η εκμάθηση μιας γλώσσας απαιτεί αρκετό χρόνο), η οποία αποθηκεύεται υποσυνείδητα, κατα την χρονοβόρα εκμάθηση της Γλώσσας.

Η “εκπαίδευση” επομένως μιας μηχανής, για την πρόβλεψη της προσωδίας του προς εκφώνηση κειμένου (και συνεπώς για την “φυσική”, ανθρώπινη εκφορά του) είναι ένα πολυσύνθετο πρόβλημα. Ας μην ξεχνάμε άλλωστε, πως για την “φυσική” εκφώνηση ενός κειμένου, εκτός των άλλων, χρειάζεται και κατανόηση του νοήματος του κειμένου.

Στα περισσότερα TTS συστήματα της εποχής μας, χρησιμοποιούνται “μηχανιστικοί” τρόποι πρόβλεψης των προσωδικών παραμέτρων, στηριζόμενοι κυρίως στην μορφή και λιγότερο στην σημασία του κειμένου, με την βοήθεια ενός συνόλου κανόνων “μορφοσυντακτικής ανάλυσης”. Η Ελληνική Γλώσσα, εξαιτίας του πλούσιου αριθμού σημείων στίξης που χρησιμοποιεί, διευκολύνει τον γρήγορο σχηματισμό προσωδικών μοντέλων, τα οποία οδηγούν σε “εύληπτα” εκφωνημένες προτάσεις.

Παρακάτω, θα παρουσιάσουμε τους αλγορίθμους που χρησιμοποιήθηκαν στον Αίσωπο, για την πρόβλεψη i) της διάρκειας των φθόγγων και ii) της βασικής συχνότητας (pitch) των φθόγγων ενός κειμένου, στηριζόμενοι σε κάποιες παραδοχές (‘μοντέλα’).

5.2.1 Μοντέλα διάρκειας (Timing)

Στον “Αίσωπο”, η διάρκεια του κάθε φθόγγου, προκύπτει από την *συνδυασμένη επίδραση* διαφόρων παραγόντων όπως η θέση του φωνήματος στην εκφώνηση¹, η ύπαρξη τόνου, η μικροπροσωδία² και το είδος των γειτονικών φθόγγων.

Ακριβώς επειδή η διάρκεια κάθε φωνήματος προκύπτει από την αθροιστική επίδραση συγκεκριμένων παραγόντων, το μοντέλο διάρκειας ονομάστηκε “Sum of Products” (S.o.P.).

5.2.1.1 Υπολογισμός μέσης τιμής και τυπικής απόκλισης της διάρκειας των φωνημάτων

Δομικό και θεμελιώδες συστατικό, του μοντέλου αυτού, είναι η μέση τιμή και η τυπική απόκλιση από τη μέση τιμή της διάρκειας όλων των φωνημάτων της Ελληνικής Γλώσσας. Στο κεφάλαιο όπου περιγράψαμε την διαδικασία παραγωγής των φωνητικών μονάδων (διφώνων), είχαμε υπογραμμίσει την εκφώνηση με σταθερό και φυσικό ρυθμό των διφώνων μέσα σε λέξεις και τις λέξεις μέσα σε πρόταση-φόρμα. Είχαμε, μάλιστα, τονίσει την σπουδαιότητα χρήσης πραγματικών λέξεων (meaningful words) έναντι του συνδυασμού γραμμάτων (logatons) στην εκφώνηση των διφώνων, έτσι ώστε να υποβοηθηθεί ο εκφωνητής στην διάρκεια ενός σταθερού ρυθμού.

Επιπλέον, η χρησιμοποίηση πρότασης στην εκφώνηση των λέξεων (και όχι η εκφώνηση μεμονωμένων λέξεων) δεν ήταν τυχαία, αφού οι μεμονωμένες λέξεις έχουν μικρότερη διάρκεια από τις ίδιες λέξεις μέσα σε προτάσεις.

Η στατιστική εκτίμηση της διάρκειας των φωνημάτων, έγινε με βάση το αρχείο που προέκυψε κατά την εξαγωγή των διφώνων από τις ηχογραφημένες προτάσεις. Στο αρχείο αυτό, για κάθε δίφωνο περιέχεται το δείγμα αρχής του αρχικού φωνήματος, το δείγμα τέλους του τελικού φωνήματος του διφώνου και το δείγμα-όριο μεταξύ των φωνημάτων. Με λίγα λόγια, για κάθε φώνημα, υπάρχουν δείγματα της διάρκειάς του, σε όλες τις φωνητικές- ακουστικές γειτονιές.

Κάνοντας την παραδοχή πως το παραπάνω δείγμα διάρκειας φωνημάτων είναι αρκετό, και συνυπολογίζοντας το γεγονός πως κατά την εξαγωγή των διφώνων, σημειώναμε μικρότερο μέρος steady-state για κάθε φώνημα (έτσι ώστε να ελαχιστοποιήσουμε τα φαινόμενα συνάρθρωσης) προσεγγίσαμε την μέση τιμή των φωνημάτων σε συνθήκες εκφώνησης ολόκληρων προτάσεων.

¹ το προφερόμενο κείμενο, μεταξύ δύο διαδοχικών παύσεων

² η εγγενής προσωδία που έχει ο κάθε φθόγγος

Για τα κροτικά (plosives), δεχτήκαμε πως αποτελούν το άθροισμα δύο ανεξάρτητων μεταβλητών, διαστήματος παύσης x_p πριν την ‘έκρηξη’ και του διαστήματος έκρηξης x_e (παραδοχή που δεν ανταποκρίνεται στην πραγματικότητα, ωστόσο είναι αναγκαία):

$$x_p + x_e \rightarrow E[x_p + x_e] = E[x_p] + E[x_e], \sigma_{x_p + x_e}^2 = \sigma_{x_p}^2 + \sigma_{x_e}^2$$

Αφού υπολογίσαμε την μέση τιμή και την τυπική απόκλιση για κάθε φθόγγο (με χρήση καθολικών συντελεστών διόρθωσης εξαιτίας του γεγονότος πως κατά την εξαγωγή των διφθώνων, σημειώναμε μικρότερο μέρος steady-state για κάθε φώνημα), χρησιμοποιήσαμε τον συνθέτη φωνής του “Αίσωπου” για μια επιπλέον διόρθωση των στατιστικών χαρακτηριστικών της διάρκειας μέσα από μια διαδικασία δοκιμής και ξαναπροσπάθειας (trial & error).

Τα τελικά αποτελέσματα, περιγράφονται από τους παρακάτω πίνακες.

Στην συνέχεια, ανα κατηγορία φθόγγων, βρέθηκε η **μεσαία τιμή (median)** τόσο για τη μέση τιμή, όσο και για την τυπική απόκλιση, και αυτές οι τιμές χρησιμοποιήθηκαν για όλους τους φθόγγους της κατηγορίας. Στους πίνακες που ακολουθούν, οι τελικές τιμές που χρησιμοποιήθηκαν στον “Αίσωπο”, εμφανίζονται με **έντονους** χαρακτήρες.

Ειδικά για το μισόφωνο /j/, πρέπει να τονίσουμε πως η διάρκεια που τελικώς επιλέχθηκε, βοηθά στην αντιμετώπιση του φαινομένου της συνίζησης, καθώς ακόμη και για τους διφθόγγους /i */ που μεταγράφονται λανθασμένα σε /j */ (όπου * οποιοδήποτε φωνηεντικό φώνημα), η διάρκεια του /j/ μπορεί να ξεγελάσει τον ακροατή και ο τελευταίος, να μην αντιληφθεί το φωνητικό σφάλμα.

Μέση τιμή και τυπική απόκλιση από τη μέση τιμή, της διάρκειας των φθόγγων της Ελληνικής (σε msecs)		
Sampa συμβολισμός φωνήματος	Μέση τιμή	Τυπική απόκλιση
φωνήεντα		
/i/	58	19
/e/	62	14
/a/	62	15
/o/	59	18
/u/	58	16
διπλά (affricates)		

/ts/	94	67
/tz/	63	29
τριβικά (fricatives)		
/f/	64	20
/v/	60	18
/T/	67	17
/D/	59	13
/s/	68	16
/z/	73	44
/x/	66	19
/G/	66	13
ρινικά (nasals)		
/m/	66	16
/n/	63	16
υγρά (liquids)		
/l/	60	14
/r/	45	16
μισόφωνα (semivowels)		
/j/	52	31
υπόλοιπα		
/C/	77	18
/jj/	66	18
/gj/	53	8

Μέση τιμή και τυπική απόκλιση
 από τη μέση τιμή,
 της **διάρκειας** των **κροτικών (plosives)** φθόγγων της Ελληνικής
 (σε msecs)

SAMPA Φωνηματικός Συμβολισμός	Μέση τιμή διαστήματος παύσης	Τυπική απόκλιση διαστήματος παύσης	Μέση τιμή διαστήματος έκρηξης	Τυπική απόκλιση διαστήματος έκρηξης	Συνολική Διάρκεια	Συνολική Τυπική Απόκλιση
--	---	---	--	--	------------------------------------	---

/p/	72	26	37	13	109	29
/b/	67	23	49	13	116	26
/t/	62	17	29	16	91	23
/d/	58	15	42	16	100	22
/k/	66	19	53	19	119	19
/g/	63	16	62	36	125	39
/c/	71	18	44	4	115	19
/gi/					53	8

5.2.1.2 Sum of Products Μοντέλο Διάρκειας (S.o.P)

Το μοντέλο διάρκειας στον Αίσωπο, περιγράφεται από τον παρακάτω τύπο:

$$du(/phoneme/) = E[] + a \sigma_x + p1 \sigma_x + p2 \sigma_x + mp \sigma_x + cf \sigma_x \text{ (msecs)},$$

όπου du είναι η τελική διάρκεια του εξεταζόμενου φωνήματος, $E[]$ είναι η μέση τιμή της διάρκειάς του, και σ_x η τυπική απόκλιση από τη μέση τιμή, όπως υπολογίστηκαν από την εξαγωγή των διφώνων (segmentation), σύμφωνα με την μεθοδολογία της προηγούμενης ενότητας. Οι υπόλοιποι συντελεστές περιγράφονται παρακάτω:

- $a=2.8$ εφόσον το φώνημα αντιστοιχεί σε τονιζόμενο φωνήεν, διαφορετικά $a=0$ (accentuation)
- $p1=1.2$ εφόσον το επόμενο φώνημα είναι μικρή παύση, διαφορετικά $p1=0$ (position)
- $p1=1.5$ εφόσον το επόμενο φώνημα είναι μεγάλη παύση, διαφορετικά $p1=0$ (position)
- $p2=1$ εφόσον το μεθεπόμενο φώνημα είναι παύση (μικρή ή μεγάλη), διαφορετικά $p2=0$ (position)
- $mp=0.6$ εφόσον το εξεταζόμενο φώνημα είναι φωνηεντικό και το επόμενο φώνημα είναι έμφωνο τριβικό (/v/, /D/, /z/, /G/), διαφορετικά $mp=0$ (microprosody)
- $cf=-1$ εφόσον ισχύουν οι παρακάτω συνδυασμοί, διαφορετικά $cf=0$ (correction factor)

Εξεταζόμενο φώνημα	Επόμενο φώνημα
/k/	/s/
/s/	/k/
/p/	/s/

/s/	/p/
/x/	/t/
/f/	/m/
/s/	/t/
/t/	/k/
/t/	/s/

Το σύνολο των παραπάνω κανόνων ενσωματώνει γενικές παρατηρήσεις σχετικά με τις μεταβολές της διάρκειας των φθόγγων, σύμφωνα με τις ανάγκες της ομιλίας. Η ύπαρξη τόνου σε μια λέξη, σημαίνει έμφαση στο συγκεκριμένο τμήμα της λέξης (συντελεστής α) και η έμφαση αυτή, μεταφράζεται σε μεγαλύτερη διάρκεια. Ο μηχανισμός έμφασης στον Αίσωπο, περιορίζεται σε τοπικό επίπεδο, στην μελέτη δηλ. των τονισμών των γραμμάτων. Ανώτερα επίπεδα επεξεργασίας σε επίπεδο σημασιολογίας (semantics), θα επέτρεπαν το υπολογισμό των φράσεων σε ένα κείμενο που χρειάζονται έμφαση, υπαγορεύοντας αντίστοιχες προσθήκες στο μοντέλο διάρκειας.

Η εγγενής προσωδία των φθόγγων λαμβάνεται υπόψιν, μέσω του συντελεστή *mr* (microprosody), στην περίπτωση των φωνηέντων που ακολουθούνται από έμφωνα τριβικά (voiced fricatives), καθώς έχει παρατηρηθεί πως η διάρκεια των φωνηέντων σ' αυτή την περίπτωση είναι λίγο μεγαλύτερη από τη μέση τιμή της (The Bell Labs Approach, 1998).

Οι παύσεις, επίσης, αυξάνουν την διάρκεια των φθόγγων που βρίσκονται στο τέλος της εκφερόμενης φράσης (συντελεστές *p1, p2*). Για τον μηχανισμό εύρεσης παύσεων και των σχηματισμό φωνητικών φράσεων, θα μιλήσουμε παρακάτω.

Τέλος, το μοντέλο διάρκειας S.o.P. διορθώνει τα λάθη πρόβλεψής του για συγκεκριμένους συνδυασμούς συμφώνων, μέσω του συντελεστή *cf* !

5.2.2 Μοντέλα Επιτόνισης (Intonation)

Μέχρι στιγμής, έχουμε εξετάσει πώς ο “Αίσωπος” μετατρέπει το κείμενο στους φθόγγους με τους οποίους θα εκφωνηθεί (Φωνηματική Μεταγραφή) και έχουμε μελετήσει τον μηχανισμό πρόβλεψης-απόδοσης τιμών διάρκειας σε κάθε φθόγγο (Μοντέλα διάρκειας-Timing).

Η Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing), θα ολοκληρωθεί με την εισαγωγή μοντέλων υπολογισμού της **βασικής συχνότητας** κάθε έμφωνου φθόγγου,

αλλά και του τρόπου με τον οποίο μεταβάλλεται, σε τοπικό επίπεδο φωνήματος και σε επίπεδο φωνητικής φράσης³ (intonation).

Οι μεταβολές της βασικής συχνότητας F_0 , της συχνότητας δηλ. με την οποία ταλαντώνονται οι φωνητικές χορδές του ομιλητή κατά την άρθρωση (articulation) των έμφωνων φθόγγων, αποτελούν τον **μηχανισμό επιτόνησης** (για τον οποίο μιλήσαμε σε προηγούμενο κεφάλαιο) και είναι υπεύθυνες για την δυνατότητα σχηματισμού ερωτηματικών ή καταφατικών προτάσεων, την έκφραση συναισθήματος και την δυνατότητα έμφασης σε συγκεκριμένα σημεία του εκφερόμενου λόγου.

Η επιτυχημένη ενσωμάτωση μεθόδων πρόβλεψης της βασικής συχνότητας των έμφωνων φωνημάτων, θα προσδώσει στην παραγόμενη ‘συνθετική’ φωνή, “τον μουσικό χρωματισμό” που κάθε Ανθρώπινη Ομιλία έχει.

5.2.2.1 Μηχανισμός εύρεσης παύσεων

Τα μοντέλα πρόβλεψης των μεταβολών της βασικής συχνότητας (intonation models) που σχηματίστηκαν, εφαρμόζονται σε επίπεδο εκφώνησης. Οι εκφωνήσεις του κειμένου αποτελούν, όπως έχουμε πει, προφερόμενο λόγο μεταξύ δύο διαδοχικών φωνητικών παύσεων. Συνεπώς, για την πλήρη εφαρμογή των μοντέλων, απαιτείται προηγουμένως, καθορισμός των φωνητικών παύσεων και προσδιορισμός των εκφωνήσεων.

Οι παύσεις υπολογίζονται με βάση τα σημεία στίξης, σύμφωνα με τον παρακάτω πίνακα:

απλή παύση διάρκεια 100 msec	διπλή παύση διάρκεια 200 msec
κόμμα (,)	
	τελεία (.)
	άνω κάτω τελεία (:)
	θαυμαστικό (!)
	ερωτηματικό (;)
παύλα (-)	

Ο καθορισμός των παύσεων και επομένως των εκφωνήσεων, με βάση τα σημεία στίξης είναι εύκολος και γρήγορος. Τί γίνεται όμως στην περίπτωση, που το κείμενο είναι φτωχό σε σημεία στίξης; Ο σχεδιασμός του “Αισωπου”, βασίστηκε εξ αρχής, στην παραδοχή ότι

³ εκφώνηση

το προς εκφώνηση κείμενο είναι συμβατό με τους **βασικούς κανόνες** της ΝεοΕλληνικής Γραμματικής, συνεπώς περιλαμβάνει σημεία στίξης.

Ωστόσο, πρόσφατα προτάθηκε αλγόριθμος εύρεσης των φωνητικών παύσεων σε κείμενο της Ελληνικής, ο οποίος δεν λαμβάνει υπόψιν του μόνο τα σημεία στίξης, αλλά προχωρά σε μορφοσυντακτική ανάλυση, απαιτώντας όμως λεξικά για κάθε μέρος του λόγου (Σουρλαντζής Ι. , 1996). Ο αλγόριθμος αυτός, δεν ενσωματώθηκε στον “Αίσωπο”, καθώς αν και αποτελεσματικός, απαιτεί κατάλογο των επιρρημάτων και των αντωνυμιών (και σε ορισμένες περιπτώσεις και των ρημάτων) της Ελληνικής, γεγονός που απαιτεί χρόνο για τον σχηματισμό των καταλόγων-λεξικών αυτών, ενώ παράλληλα καθυστερεί την επεξεργασία κατά την εκτέλεση του προγράμματος, λόγω της αναζήτησης για κάθε λέξη σε μεγάλα λεξικά.

Για λόγους πληρότητας, θα παρουσιάσουμε εν συντομία το αλγόριθμο αυτό.

Οι λέξεις χωρίζονται σε 3 κατηγορίες, ανάλογα με το αν μπορούν να ενωθούν με τις γειτονικές τους στην ίδια εκφώνηση:

α. λέξεις που μπορούν να συνενωθούν με την επόμενη

- άρθρα
- αντωνυμίες, εκτός από τις κτητικές και τις ερωτηματικές
- συμπλεκτικοί, διαζευκτικοί,, αντιθετικοί, ειδικοί, χρονικοί, αιτιολογικοί, αποτελεσματικοί, διστακτικοί σύνδεσμοι
- τα μόρια <<ας, θα, να, μα, για>>
- οι προθέσεις
- τα επιρρήματα, εκτός από τα ερωτηματικά

β. λέξεις που μπορούν να ενωθούν με την προηγούμενη

Οι κτητικές αντωνυμίες << μου, σου, του, της, μας, σας, τους >>

γ. λέξεις που δεν ενώνονται ούτε με την επόμενη ούτε με την προηγούμενη

Λέξεις της δεύτερης κατηγορίας εμφανίζονται και στην πρώτη. Οι κτητικές αντωνυμίες << μου, σου του >> λειτουργούν ως προσωπικές προστά από τα ρήματα, συνεπώς για την επιτυχή λειτουργία του αλγορίθμου χρειάζεται κατάλογος των ρημάτων. Βεβαίως, η δεύτερη κατηγορία λέξεων μπορεί να αγνοηθεί, εκτός από την περίπτωση όπου εμφανίζονται κτητικές και προσωπικές αντωνυμίες διαδοχικά π.χ. η φίλη **μου μου** είπε στο **αυτί...**

Τα όρια της κάθε εκφώνησης καθορίζονται εκτός από τα σημεία στίξης και από το είδος των λέξεων, με βάση τον παραπάνω διαχωρισμό: κάθε λέξη ανήκει στην εκφώνηση της προηγούμενης λέξης, εφόσον ανήκει στην κατηγορία (β) ή η προηγούμενη λέξη ανήκει στην κατηγορία (α), διαφορετικά, υπάρχει παύση και η λέξη ξεκινά καινούργια εκφώνηση.

5.2.2.2 Linear Intonation models στον “Αίσωπο”

Με βάση τις παύσεις, τα φωνήματα χωρίζονται σε εκφωνήσεις, και για κάθε εκφώνηση αποθηκεύεται:

- i. το πλήθος των φωνημάτων που περιέχει.
- ii. εάν είναι ερωτηματική ή όχι.
- iii. εάν είναι αρχική, ενδιάμεση ή τελική εκφώνηση στην πρόταση.

Στην συνέχεια, για κάθε εκφώνηση υπολογίζεται η “ μέση καμπύλη τονικότητας” , με καθαρά γραμμικό τρόπο, με βάση τον αριθμό των φωνημάτων που περιέχει, και με βάση την αρχική (start pitch) και την τελική τιμή του pitch (end pitch), ανάλογα με το είδος της:

Αρχικό και τελικό pitch, για κάθε είδος εκφώνησης					
Είδος Εκφώνησης στην Πρόταση					
	αρχική	εσωτερική	ερωτηματική	κατάληξη σε απλή παύση	κατάληξη σε διπλή παύση
start pitch (Hz)	107	105			
end pitch (Hz)			112	100	95

Η μέση καμπύλη τονικότητας, για κάθε εκφώνηση προκύπτει από τον παρακάτω τύπο:

$$\text{pitch} = \text{start pitch} + i (\text{end pitch} - \text{start pitch}) / \text{plithos},$$

όπου plithos είναι το πλήθος των φωνημάτων της εκφώνησης, και i είναι η θέση του φωνήματος, του οποίου το pitch θέλουμε να προβλέψουμε.

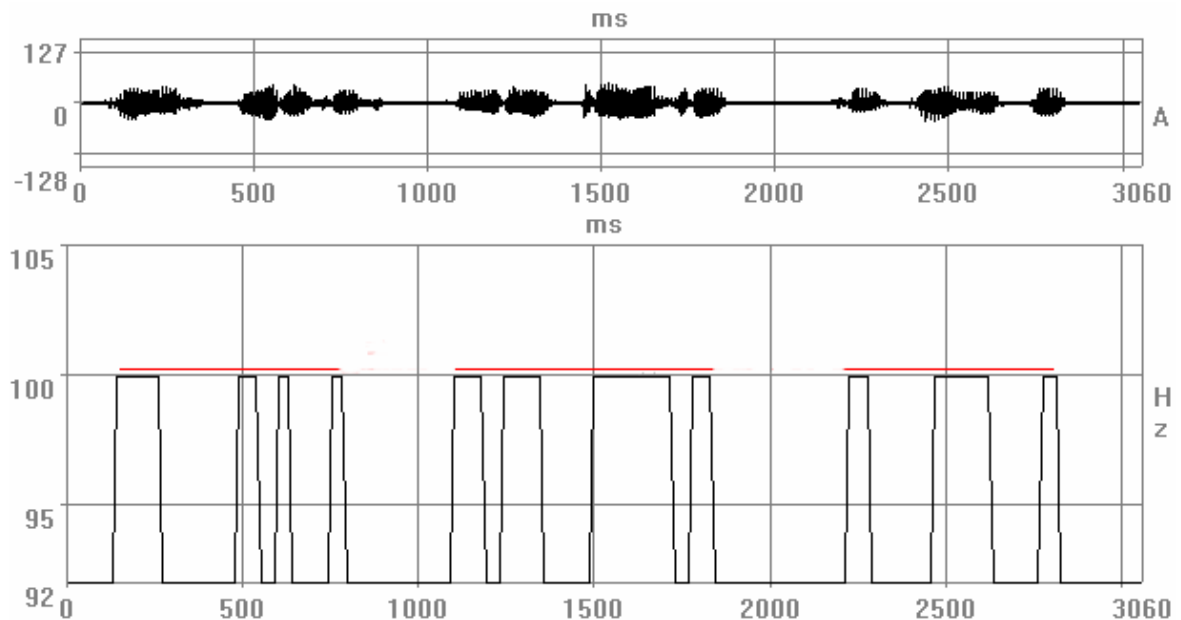
Για την περίπτωση που η εκφώνηση είναι μικρή ($\text{plithos} \leq 12$) το start pitch=106 για αρχική εκφώνηση, ενώ το end pitch=114 για ερωτηματική εκφώνηση.

Ο παραπάνω πίνακας, συνοψίζει γενικές παρατηρήσεις σχετικά με τις μεταβολές του pitch, όπως την μείωση του pitch στο τέλος μιας φράσης ή μιας πρότασης, το ανέβασμα του pitch σε μια ερώτηση, ενώ μέσω της ανακατασκευής της μέσης τονικής καμπύλης για κάθε εκφώνηση, δίνεται *έμφαση στην αντίστοιχη φράση*, ενώ παράλληλα πετυχαίνεται η

“τονική ανανέωση” της πρότασης, με συνέπεια μια εύληπτη, ευχάριστη “ακουστικά” συνθετική φωνή.

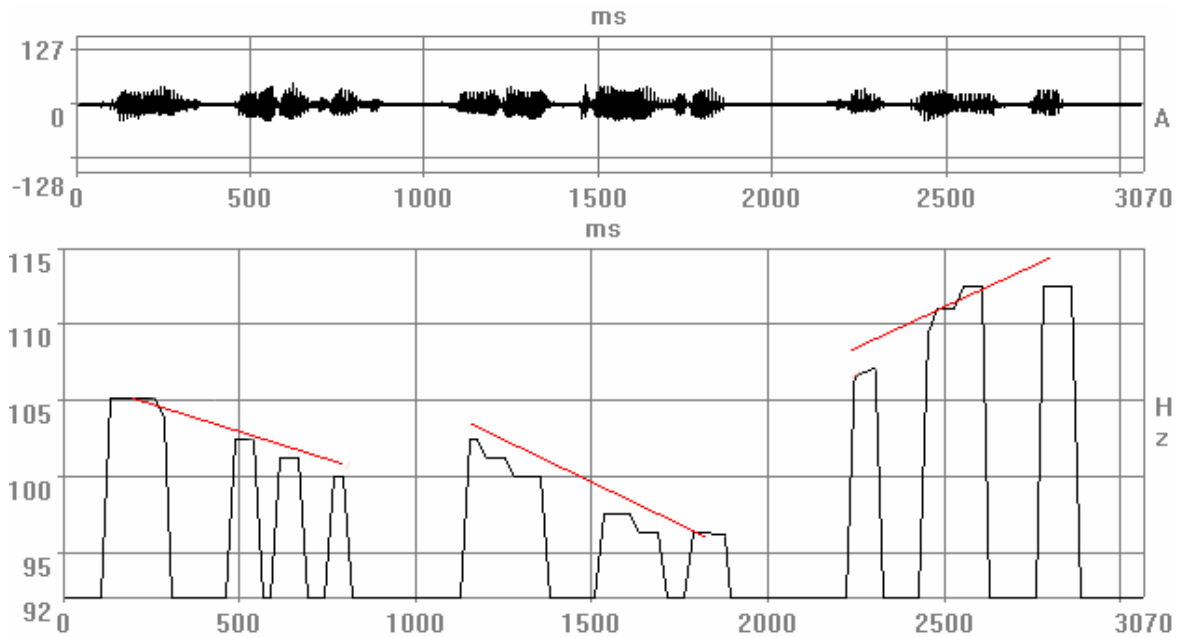
Στα παρακάτω σχήματα, φαίνεται η εφαρμογή των παραπάνω, στην φράση <<Καλησπέρα σας, κύριε πρόεδρε. Τί κάνετε ;>> η οποία, θα χωριστεί κατά την επεξεργασία, σε 3 εκφωνήσεις:

- i. /kalisperasas/ → αρχική στην πρόταση, καταφατική που τελειώνει σε απλή παύση → start pitch=107 Hz, end pitch= 100 Hz.
- ii. /kirieproeDre/ → εσωτερική στην πρόταση, καταφατική που τελειώνει σε διπλή παύση → start pitch=105 Hz, end pitch=95 Hz.
- iii. /tikanete/ → αρχική, ερωτηματική → start pitch=107 Hz, end pitch=112 Hz



[Η μεταβολή της βασικής συχνότητας, στην φράση << Καλησπέρα σας, κύριε πρόεδρε. Τί κάνετε; >> συναρτήσει του χρόνου.

Οι 3 εκφωνήσεις της φράσης έχουν εκφωνηθεί από τον “Αίσωπο” με σταθερό pitch=100 Hz, και διάρκειες φωνημάτων υπολογισμένες από το S.o.P. μοντέλο διαρκειών. Η συνθετική φωνή, είναι μονότονη και επομένως κουραστική, χωρίς την δυνατότητα έκφρασης ερώτησης]



[Η μεταβολή της βασικής συχνότητας συναρτήσει του χρόνου για την ίδια φράση, εκφωνημένη με το γραμμικό μοντέλο επιτόνισης του “Αίσωπου”.

Οι 3 εκφωνήσεις, έχουν μια μέση τονική καμπύλη, με διαφορετική κλίση εξαρτώμενη από το είδος της εκφώνησης. Η παραγόμενη συνθετική φωνή, είναι ευχάριστα καταληπτή, ενώ η ερώτηση στο τέλος του κειμένου γίνεται αντιληπτή]

Το παραπάνω γραμμικό μοντέλο επιτόνισης (intonation model), ονομάστηκε Linear Intonational Model (LIM). Ωστόσο, το μοντέλο αυτό, δεν λαμβάνει υπόψιν τους τοπικούς τονισμούς των λέξεων, οι οποίοι είναι ευδιάκριτοι σ’ ένα κείμενο εξαιτίας του αντίστοιχου τονικού σημαδιού, το οποίο ο Ελληνικός Γραπτός Λόγος “απλόχερα” προσφέρει, σε αντίθεση με άλλες Γλώσσες όπου τονικά σημάδια λέξεων δεν υπάρχουν (π.χ. στα Αγγλικά).

Η ενσωμάτωση της επιτόνισης (μεταβολή του τόνου, (pitch)) σε επίπεδο λέξης, με βάση το τονικό της σημάδι, οδήγησε στην εξέλιξη του Γραμμικού μοντέλου Επιτόνισης (LIM) στο Γραμμικό μοντέλο Υπέρθεσης (Linear Superpositional Model, (LSM)).

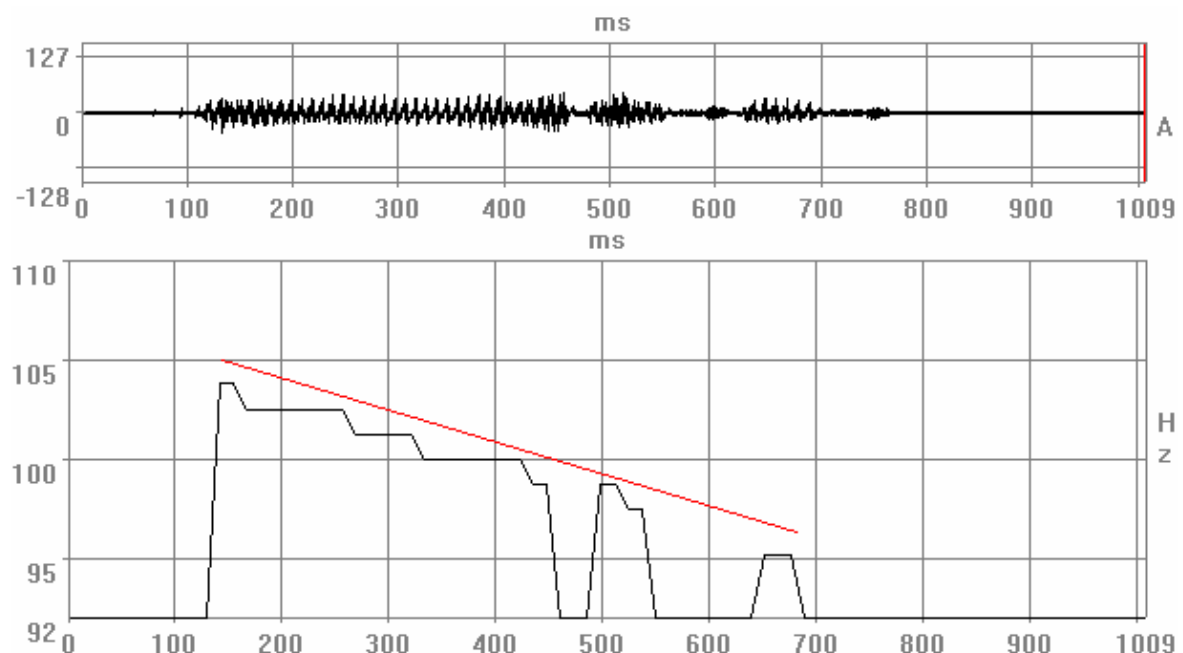
Ουσιαστικά, στο μοντέλο αυτό, η μέση τονική καμπύλη για κάθε εκφώνηση διατηρείται, ενώ το pitch αυξάνεται κατά 5 Hz, για το 70% της διάρκειας του φωνηεντικού φωνήματος, της λέξης που τονίζεται.

Το Γραμμικό Μοντέλο Υπέρθεσης, αποτελεί το προκαθορισμένο (default) μοντέλο που χρησιμοποιεί ο “Αίσωπος”. Είναι ένα απλό και γρήγορο μοντέλο, που προέκυψε από το “πείραμα της εφημερίδας”. Όταν διαβάζουμε μια στήλη σε μια εφημερίδα, διαβάζουμε “διερευνητικά” αξιοποιώντας τα σημεία στίξης, καθώς δεν ξέρουμε τί ακολουθεί. Διαβάζουμε δηλαδή, αξιοποιώντας τα τοπικά γλωσσικά στοιχεία. Εάν διαβάσουμε

φωναχτά , μια στήλη σε μια εφημερίδα, κρύβοντας τις επόμενες γραμμές, θα δούμε πως αν και μας ενοχλεί το γεγονός ότι δεν μπορούμε να δούμε τα επόμενα γλωσσικά στοιχεία, εντούτοις είμαστε ικανοί να διαβάσουμε ικανοποιητικά το κείμενο, αξιοποιώντας μόνο τα τοπικά γλωσσικά στοιχεία και **κυρίως τα τοπικά σημεία στίξης**⁴.

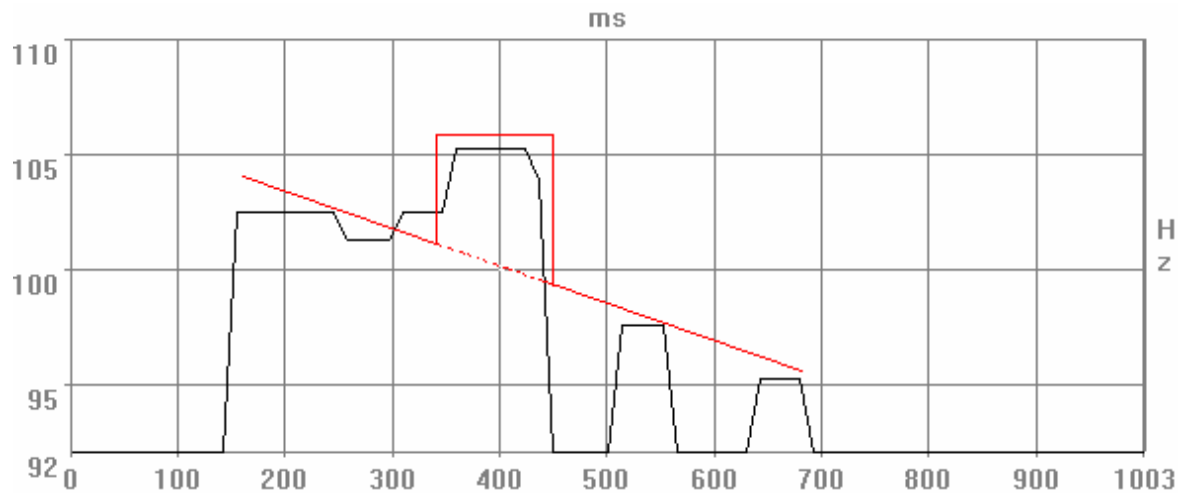
Ωστόσο, το μοντέλο επιτόνισης στον “Αίσωπο” (LSM), δεν είναι μοντέλο που στηρίζεται μόνο σε τοπικά στοιχεία. Έχει προηγηθεί ο μηχανισμός εύρεσης παύσεων και το κείμενο έχει χωριστεί σε εκφωνήσεις.

Συμπερασματικά, θα λέγαμε πως το γραμμικό μοντέλο υπέρθεσης στον “Αίσωπο” είναι ένα μοντέλο που δανείζεται στοιχεία, **τόσο** από τα **τονικά μοντέλα επιτόνισης**, που θεωρούν πως οι καμπύλες μεταβολής της βασικής συχνότητας παράγονται από σειρά ανεξάρτητων διακεκριμένων τόνων, οι οποίοι υπολογίζονται σε τοπικό επίπεδο (tone sequence models), **όσο** και από τα **ιεραρχικά μοντέλα επιτόνισης**, που θεωρούν πως οι καμπύλες βασικής συχνότητας παράγονται από συνδυασμένη επίδραση (superposition) παραγόντων, σε επίπεδο συλλαβής λέξης, φράσης, πρότασης ακόμη και παραγράφου (hierarchical superposition models).

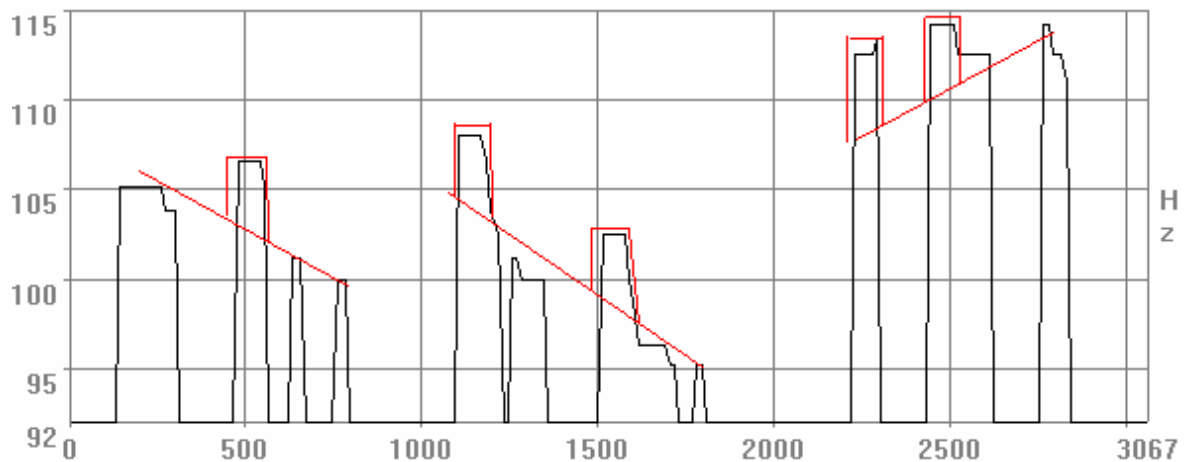
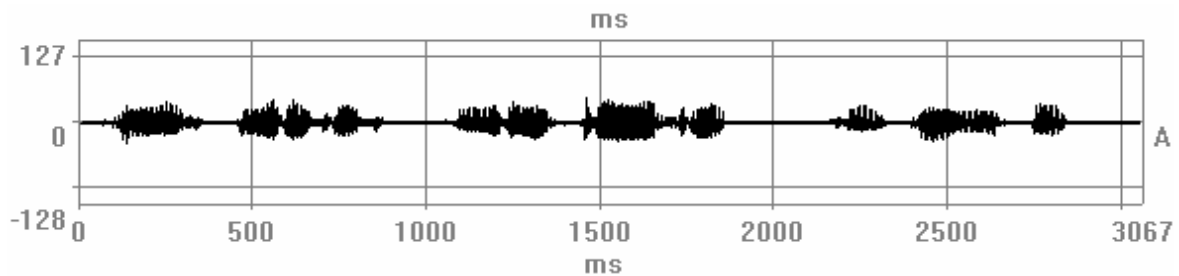


[Η μεταβολή της βασικής συχνότητας για την φράση <<καλημέρα σας>>, η οποία εκφωνήθηκε σύμφωνα με το μοντέλο L.I.M. Ο τονισμός του γράμματος <ε>, δεν λαμβάνεται υπόψιν από τον μηχανισμό επιτόνισης, αν και λαμβάνεται υπόψιν στον καθορισμό της διάρκειας (μοντέλο S.o.P.)]

⁴ Βεβαίως, εάν ξέρουμε τί ακολουθεί, ή εάν έχουμε ξαναδιαβάσει το κείμενο, τότε μπορούμε να το



[Η μεταβολή της βασικής συχνότητας συναρτήσει του χρόνου στην φράση <<καλημέρα σας>>, σύμφωνα με το μοντέλο επιτόνισης L.S.M. Τα πλεονεκτήματα του μοντέλου L.I.M. διατηρούνται, ενώ λαμβάνεται υπόψιν ο τοπικός τονισμός της λέξης, με αύξηση κατά 5 Hz της βασικής συχνότητας του τονιζόμενου φωνηεντικού φωνήματος /e/ της λέξης <<καλημέρα>>]



[η καμπύλη μεταβολής της βασικής συχνότητας F_0 , στην φράση <<Καλησπέρα σας, κύριε πρόεδρε. Τι κάνετε;>> εκφωνημένη από τον “Αίσωπο”, με το μοντέλο διαρκειών S.o.P. και το μοντέλο επιτόνισης

L.S.M. Φαίνεται, για κάθε εκφώνηση, η μέση τονική καμπύλη, ενώ ευδιάκριτη είναι η επιπλέον επιτόνιση στα τονιζόμενα φωνήεντα]

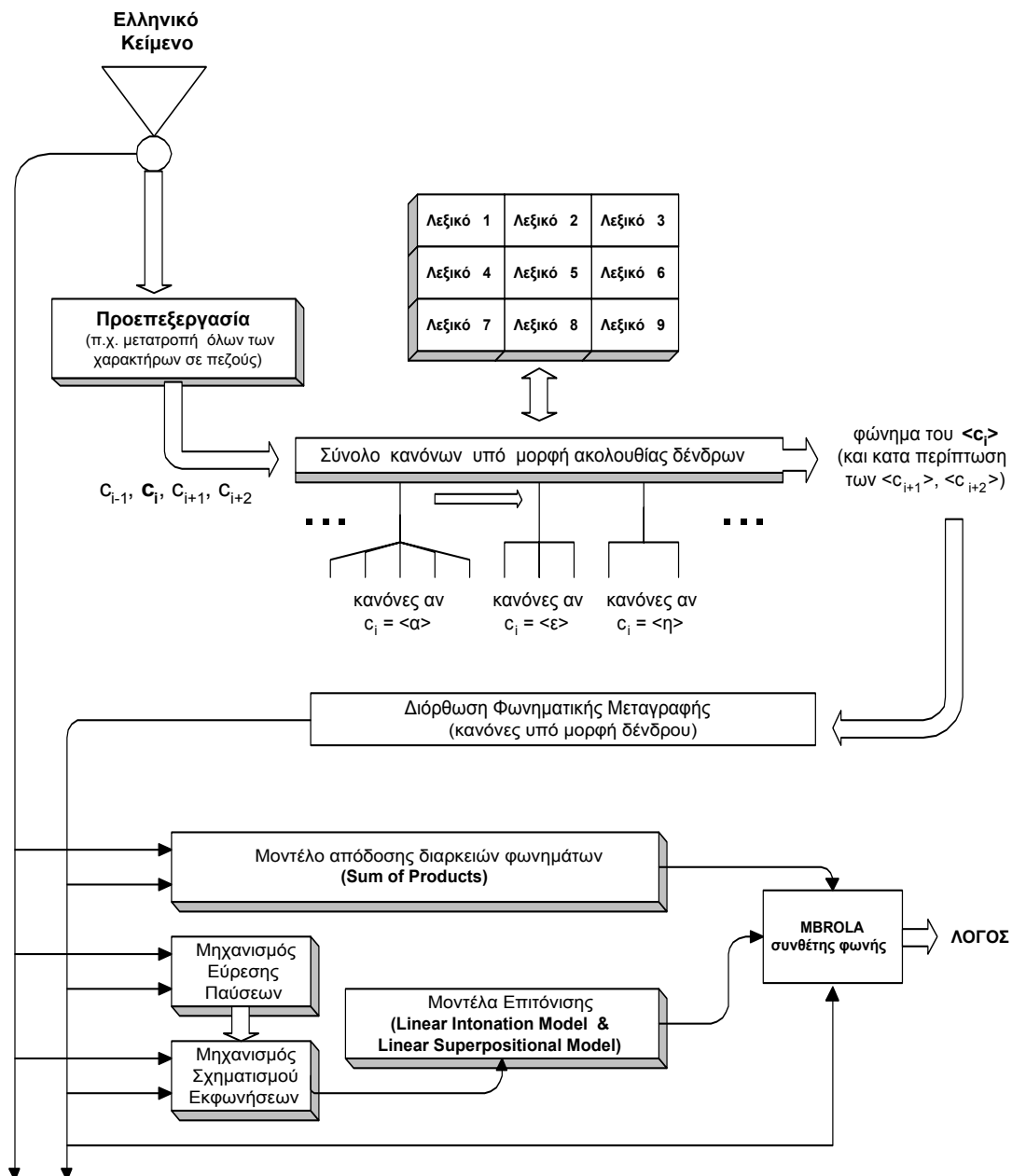
Κεφάλαιο έξι

Επίλογος

Στο κεφάλαιο αυτό, θα συνοψίσουμε τα λειτουργικά μέρη στο ESoPoS-TTS, θα εκτιμήσουμε την καταληπτότητα αλλά και τη γενικότερη “ποιότητα” της παραγόμενης συνθετικής φωνής και θα προτείνουμε μελλοντικές βελτιώσεις.

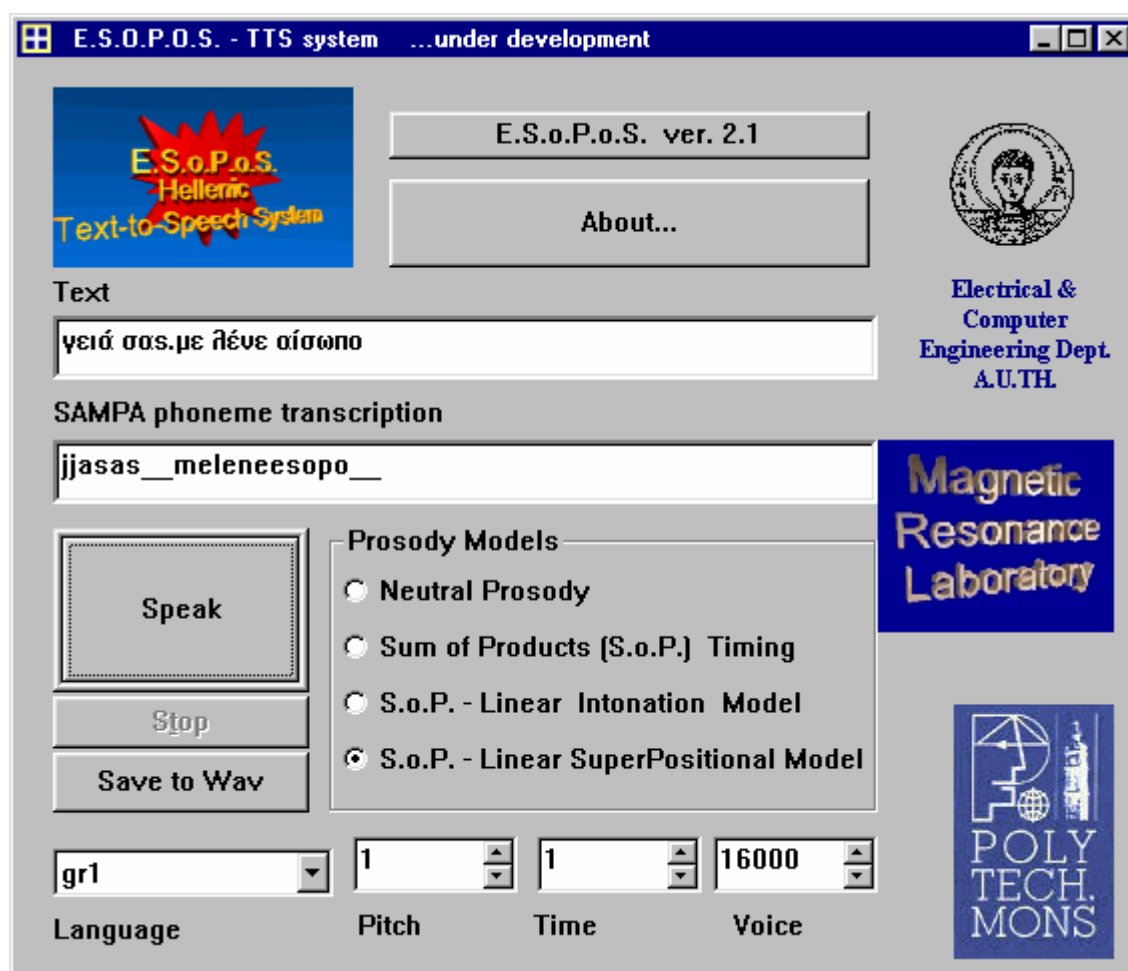
6.1 Η τελική δομή

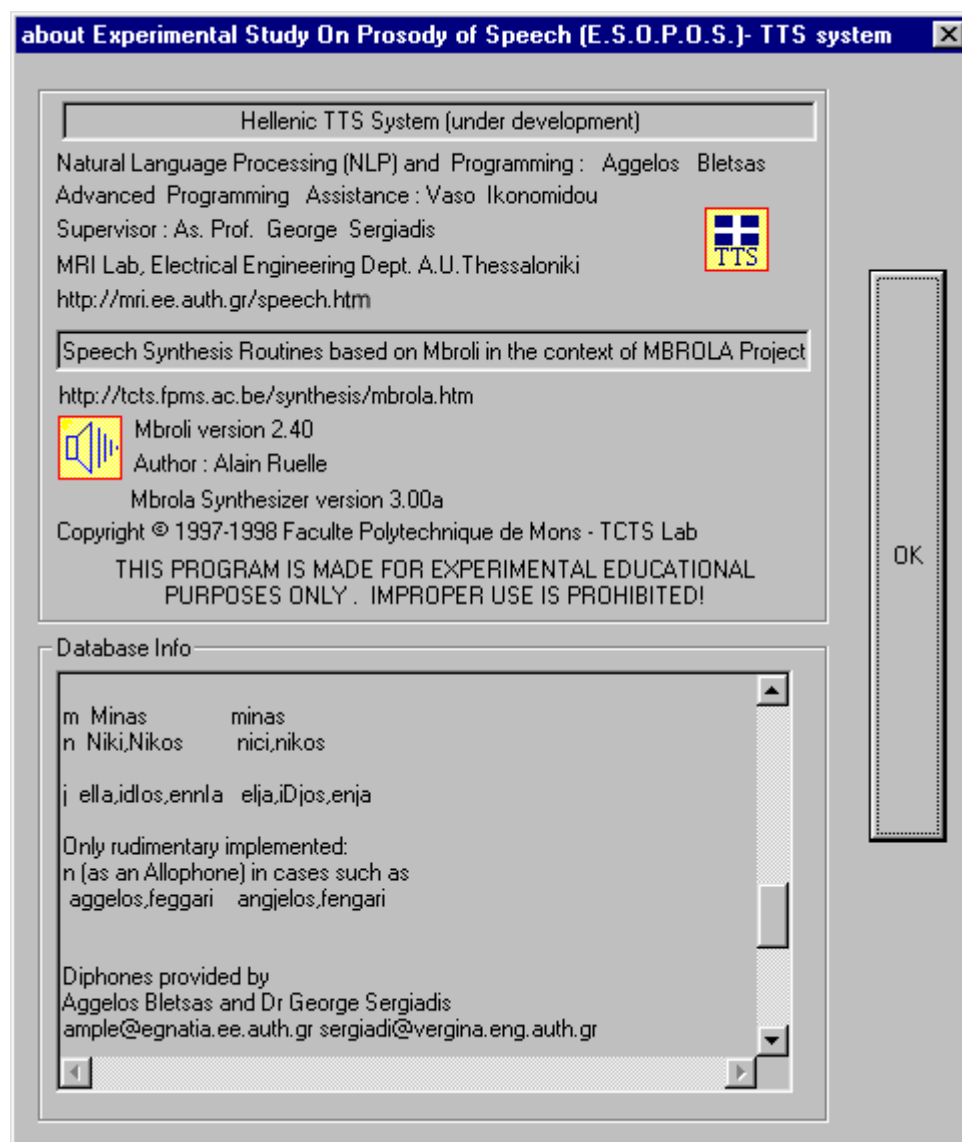
Οι λειτουργίες που περιγράψαμε αναλυτικά στα προηγούμενα κεφάλαια, συνοψίζονται στο παρακάτω διάγραμμα:



Στο παραπάνω διάγραμμα φαίνεται καθαρά, πως το κείμενο εισόδου χρησιμοποιείται σε όλα τα στάδια επεξεργασίας, και όχι μόνο στο αρχικό. Ο παραπάνω αλγόριθμος, υλοποιήθηκε σε Visual C++ 5, της οποίας ο αντικειμενοστραφής χαρακτήρας υποβοήθησε στην δημιουργία μιας πλατφόρμας εργασίας με "ανοικτή αρχιτεκτονική", φιλικής σε κάθε προγραμματιστή, ο οποίος μελλοντικά θα ενδιαφερθεί στην εξέλιξη του συγκεκριμένου TTS.

Η τελική μορφή του εκτελέσιμου προγράμματος (dialog based), φαίνεται παρακάτω:





6.2 Αξιολόγηση

Για την αξιολόγηση της “ποιότητας” της παραγόμενης συνθετικής φωνής απαιτείται η διενέργεια καθολικά αποδεκτών τεστ, τα οποία θα εξετάσουν:

- την καταληπτότητα της συνθετικής φωνής (**intelligibility**)
- την κατανόηση της συνθετικής φωνής (**comprehension**)
- την φυσικότητα της συνθετικής φωνής (**naturalness**)

Δυστυχώς, αν και η έννοια της “ποιότητας” είναι a priori υποκειμενική, δεν υπάρχουν κοινώς αποδεκτά τεστ από την επιστημονική κοινότητα, για τη μέτρηση των παραπάνω χαρακτηριστικών στη συνθετική φωνή, την οποία παράγει ένα Text-to-Speech σύστημα.

Ωστόσο, για λόγους πληρότητας της εργασίας αυτής, προχωρήσαμε στον σχηματισμό συγκεκριμένου τεστ, για την μέτρηση της καταληπτότητας της συνθετικής φωνής, και

ελπίζουμε το τεστ αυτό, να αποτελέσει το standard στην εξέταση της καταληπτότητας της παραγόμενης φωνής, οποιουδήποτε Ελληνικού Συνθέτη.

Το τεστ καταληπτότητας που σχηματίστηκε, είναι ένα Modified Rhyme Test, στο οποίο εξετάζεται περισσότερο, η καταληπτότητα των συμφώνων καθώς αυτά είναι τα πιο “δύσκολα” στη Σύνθεση Φωνής.

Συγκεκριμένα, σχηματίστηκαν 30 εξάδες ομόηχων λέξεων, οι οποίες διέφεραν μόνο σε κάποιο συμφωνικό φώνημα, το οποίο βρισκόταν σε C-V (Consonant-Vowel) ή C-V-C (Consonant-Vowel-Consonant) συνδυασμό, αποφεύχθηκαν δηλ. οι ομόηχες λέξεις όπου το υπό εξέταση συμφωνικό φώνημα, βρισκόταν σε σύμπλεγμα συμφώνων.¹

Για να “δυσκολέψουμε” το τεστ, στην πλειοψηφία των λέξεων που εκφωνήθηκαν (16/30), το εξεταζόμενο σύμφωνο ήταν κροτικό (plosive), καθώς αυτή η κατηγορία συμφώνων είναι η πιο δύσκολη στη Σύνθεση.

Οι 30 λέξεις εκφωνήθηκαν σε ομάδα 10 “μη πεπειραμένων” ακροατών, στο εργαστήριο Μαγνητικού Συντονισμού της σχολής μας, κάτω από όχι ιδανικές συνθήκες, λόγω θορύβου των παρακείμενων δρόμων. Οι ακροατές ζητήθηκαν να επιλέξουν τη λέξη που ταιριάζει περισσότερο στη λέξη που θα ακούσουν, από κάθε εξάδα λέξεων.

Το μέσο σφάλμα, κάτω από τις συνθήκες που περιγράψαμε, ήταν 11% (➔ ποσοστό επιτυχίας 89%) το οποίο κρίνεται ιδιαίτερα ενθαρρυντικό. Τα σφάλματα επικεντρώθηκαν στο plosive /t/, εξαιτίας της μάλλον λάθος διάρκειάς του.

Modified Rhyme Test για τα Ελληνικά

μέσο σφάλμα στο	11%
ESoPoS-TTS	
φυσική φωνή	0.8%

Το MRT που εφαρμόσαμε, δεν μοιάζει με τα αντίστοιχα τεστ που εφαρμόστηκαν για τα υπόλοιπα TTS συστήματα της Αγγλικής (π.χ. άγνωστο είναι το ποσοστό των plosives που εξετάστηκαν, όταν στον “Αίσωπο” ήταν 16/30), συνεπώς το “σκορ” αυτό δεν είναι άμεσα συγκρίσιμο με το μέσο σφάλμα των TTS της Αγγλικής, όπως παρουσιάζεται στη σχετική βιβλιογραφία.

Εντούτοις, για λόγους πληρότητας παρουσιάζουμε το μέσο σφάλμα καταληπτότητας για τα πιο γνωστά TTS συστήματα της Αγγλικής.

Modified Rhyme Test για τα Αγγλικά

DECtalk	5.72%
Prose 3.0	5.72%
MITalk-79	7.00%
Amiga	12.25%
Infovox	12.50%
TSI-Proto 1	17.75%
Smoothtalker	27.22%
Votrax Type'n'Talk	27.44%
Echo	35.56%

[Logan, 1989]

Για την εξέταση της κατανόησης (comprehension), ο “Αίσωπος” εκφώνησε κείμενο και οι ακροατές στην συνέχεια, απάντησαν σε ερωτήσεις σχετικά με το νόημα του κειμένου. Μεταξύ των κειμένων που εκφωνήθηκαν, ήταν και λογοτεχνικά ποιήματα. Η κατανόηση των κειμένων ήταν, σε γενικές γραμμές πλήρης, χωρίς σοβαρά προβλήματα.

Στην εξέταση της φυσικότητας (naturalness), εκτιμάται η γενικότερη ποιότητα του TTS, καλώντας τους ακροατές να διαλέξουν μεταξύ των ίδιων φράσεων εκφωνημένων από διαφορετικά TTS συστήματα, με συνέπεια τον σχηματισμό Mean Opinion Scores (MOS). Μέχρι τη στιγμή συγγραφής της εργασίας αυτής, οι υπόλοιποι συνθέτες φωνής της Ελληνικής, ήταν υποδεέστεροι του Συνθέτη Φωνής στον “Αίσωπο”, όπως μπορεί ο καθένας να εκτιμήσει, με βάση τις κυματομορφές-παραδείγματα του κάθε συστήματος, όπως βρίσκονται στο Διαδίκτυο. Συνεπώς, ο σχηματισμός Mean Opinion Scores, ήταν άνευ νοήματος.

Για τις υπόλοιπες δυνατότητες του Συστήματος, θα λέγαμε πως το ποσοστό σωστής φωνηματικής μεταγραφής, έφτασε το 92% με το ποσοστό σφάλματος να οφείλεται στο φαινόμενο της συνίζησης, και να είναι αντιστρόφως ανάλογο της “ενημέρωσης” των λεξικών μορφημάτων-εξαιρέσεων, υπεύθυνων για τον χειρισμό των καταχρηστικών διφθόγγων.

Τέλος, τα μοντέλα Προσωδίας (SoP, LIM, LSM), αποτέλεσαν μια ικανοποιητική πρώτη προσέγγιση, συνεισφέροντας σε μια όχι κουραστική, ούτε μονότονη συνθετική φωνή.

¹ Η πλήρης λίστα των λέξεων στο Modified Rhyme Test, παρουσιάζεται στο παράρτημα

Η βάση των διφώνων, απαιτεί 1.8 MB, η επεξεργασία του προγράμματος είναι ταχύτατη ενώ η χρησιμοποίηση δυναμικών μέσων αποθήκευσης (λίστες και όχι πίνακες), επιτρέπει την επεξεργασία κειμένου “απεριόριστου” μήκους.

Ο ESoPos-TTS, επιδείχθηκε στα πλαίσια της Infosystem '98, στο εκθεσιακό χώρο της διεθνούς Εκθέσεως Θεσσαλονίκης, και απέσπασε τον θαυμασμό της επιστημονικής κοινότητας, τον ενδιαφέρον του επιχειρηματικού κόσμου και τα επαινετικά σχόλια του απλού κόσμου.

6.3 Μελλοντικές Βελτιώσεις

Ο E.S.o.P.o.S. αποτέλεσε την αφορμή της πρώτης επαφής μας, με τον χώρο των Γλωσσικών Τεχνολογιών. Η υπάρχουσα Τεχνογνωσία στον χώρο αυτό, προέρχεται από το εξωτερικό και συνεπώς εξειδικεύεται στις ιδιαιτερότητες Γλωσσών διαφορετικών από την Ελληνική.

Η μεθοδολογία που ακολουθήσαμε στην εργασία αυτή, ήταν η αξιολόγηση των πολύχρονων ερευνητικών προσπαθειών στον τομέα των ξένων Text-to-Speech Συστημάτων, η προσαρμογή των πετυχημένων τεχνικών στην Ελληνική Γλώσσα και η ενσωμάτωση τους, στην σχεδίαση ενός Ελληνικού Συστήματος TTS. Άλλωστε, η σχετική Ελληνική βιβλιογραφία, ήταν ιδιαίτερα φτωχή.

Ο E.S.o.P.o.S. αποτελεί μια πρώτη προσέγγιση στο πρόβλημα που λέγεται “Αυτόματη Εκφώνηση Κειμένου”, καλύπτοντας ταυτόχρονα το κενό Τεχνογνωσίας στον Ελληνικό χώρο. Επομένως, θα πρέπει να κριθεί ως μια πρώτη, πειραματική προσπάθεια η οποία δίνει θαυμαστά αποτελέσματα, ωστόσο επιδέχεται βελτιώσεων:

- **φωνηματική μεταγραφή**

Θα πρέπει να συμπεριληφθεί χειρισμός των συντομογραφιών (π.χ. , κ.λ.π. , π.Χ. ...), των ειδικών συμβόλων (% , \$, £ ...) και αριθμών (ακέραιων ή δεκαδικών) στο σωστό γένος και πτώση. Η παρατήρηση των καταλήξεων των γειτονικών λέξεων θα δώσει πληροφορίες για το κατάλληλο γένος και πτώση.

Η περαιτέρω ενημέρωση των λεξικών μορφημάτων-εξαιρέσεων, θα βοηθήσει στην πιο πετυχημένη αντιμετώπιση του φαινομένου της συνίζησης, συντελώντας σε πιο ακριβή φωνηματική μεταγραφή.

- **προσωδία**

Ο μηχανισμός εύρεσης παύσεων θα πρέπει να βελτιωθεί και να συμπληρωθεί με ένα σύνολο κανόνων που θα εκμεταλεύεται λέξεις-κλειδιά (function words), όπως τα άρθρα, οι σύνδεσμοι και οι αντωνυμίες, των οποίων ο αριθμός είναι σχετικά μικρός (σε αντίθεση

με τον αριθμό των επιρρημάτων) και η μορφή δεν αλλάζει, ανεξάρτητα με την ανανέωση της Γλώσσας. Η αναγνώρισή τους στην φράση θα γίνεται με χρήση λεξικού (γι' αυτό ο αριθμός τους πρέπει να είναι μικρός).

Η βελτίωση του μηχανισμού εύρεσης παύσεων, θα βελτιώσει τον χωρισμό του κειμένου σε φωνητικές φράσεις (εκφωνήσεις) βελτιώνοντας και τον ρυθμό του παραγόμενου λόγου. Ειδικά για τις διάρκειες των φωνημάτων, θα λέγαμε πως πρέπει να προσαρμόζονται και στο μήκος όχι μόνο της εκφώνησης αλλά ολόκληρου του κειμένου, καθώς τα 'μεγάλα' κείμενα έχουμε την τάση να τα προφέρουμε γρηγορότερα. Στο μοντέλο διάρκειας, επομένως, θα πρέπει να υπάρχει και ένας συντελεστής

$r * \sigma(\text{phoneme})$ όπου $r < 0$ και $|r|$ ανάλογο του μήκους του κειμένου.

Επίσης, θα πρέπει να επανεξεταστεί η διάρκεια του /t/ (κυρίως του διαστήματος σιωπής που προηγείται της έκρηξης κατά την άρθρωση του /t/), καθώς το συγκεκριμένο φώνημα εμφάνισε τα περισσότερα σφάλματα καταληπτότητας στα MRT.

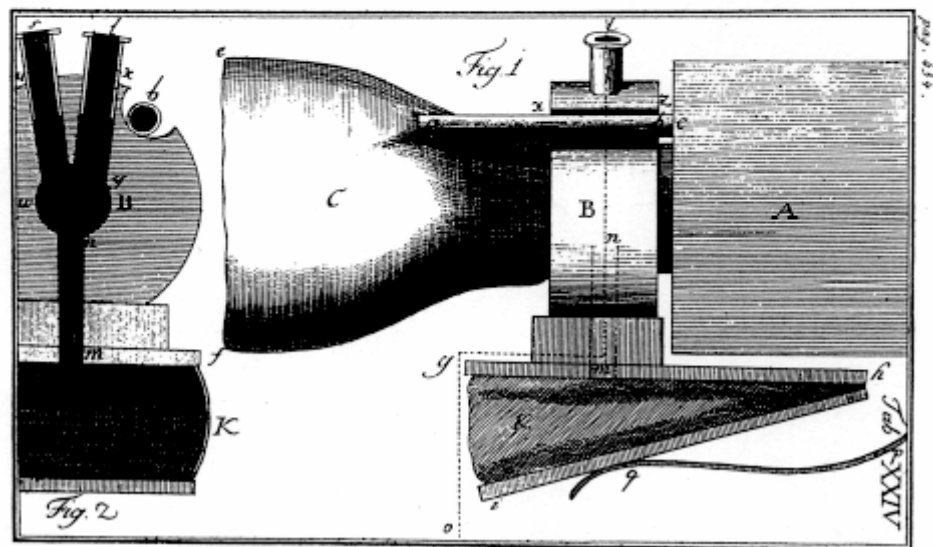
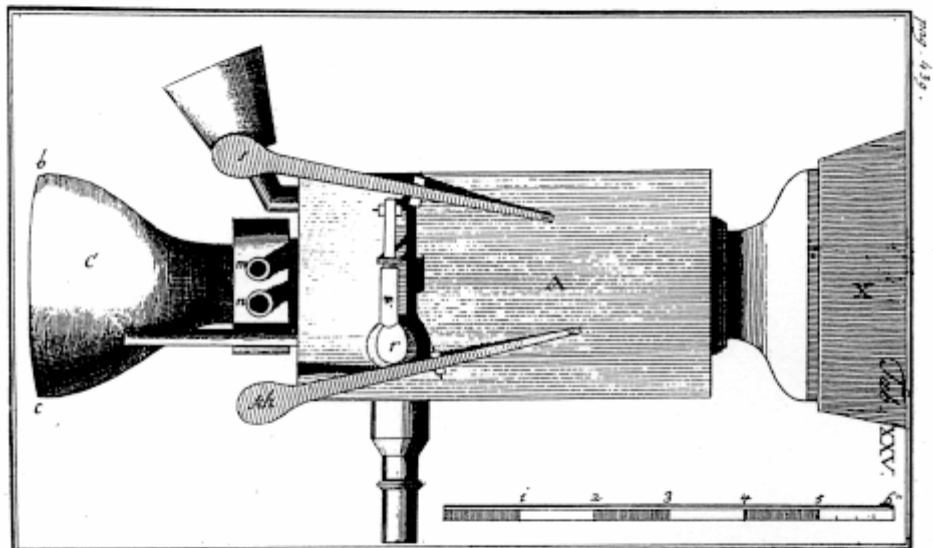
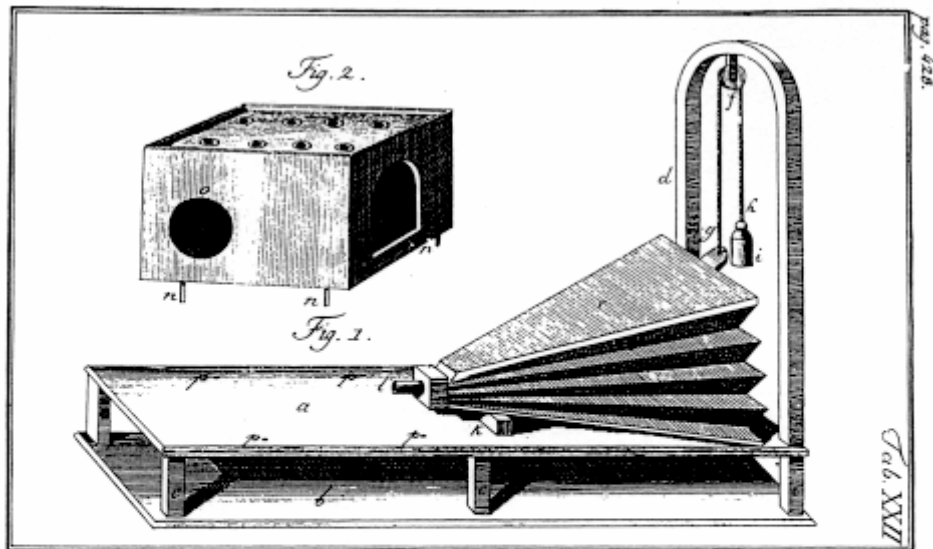
Για τα μοντέλα επιτόνισης θα πρέπει να δοκιμαστούν και μη γραμμικές μέσες τονικές καμπύλες (π.χ. παραβολικές), των οποίων τα χαρακτηριστικά θα μετρηθούν από 'φυσική' φωνή.

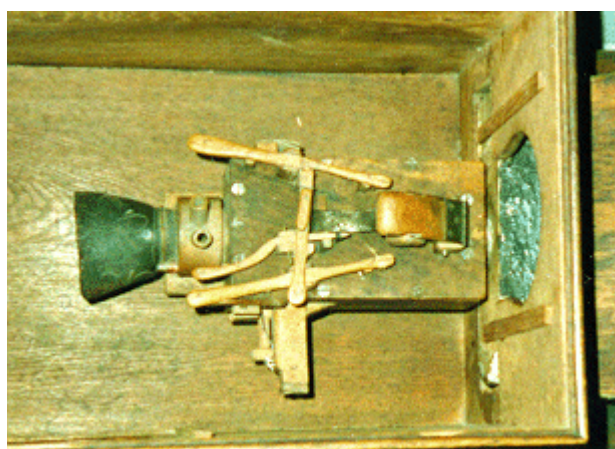
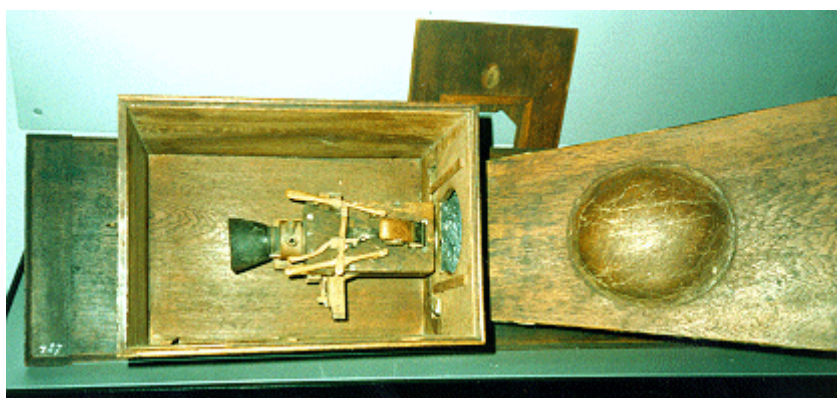
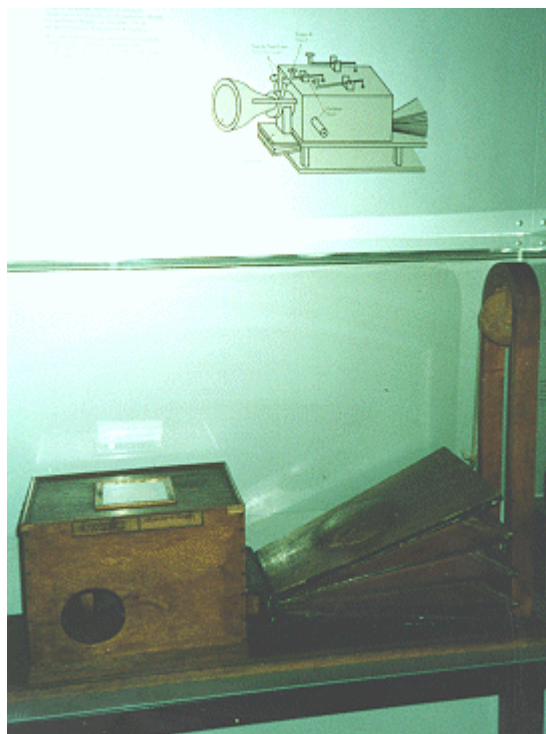
Ειδικά για τις ερωτηματικές εκφωνήσεις, θα πρέπει να γίνει ο διαχωρισμός εάν είναι επιβεβαιωτικές ή όχι, ενδεχομένως με ανίχνευση του είδους των αντωνυμιών, και να προβλεφθούν μέσες τονικές καμπύλες, κατάλληλες για κάθε περίπτωση.

Τέλος, θα πρέπει να επισκοπηθούν και στοχαστικές διαδικασίες επεξεργασίας Φυσικής Γλώσσας (νευρωνικά δίκτυα, Hidden Markov Models), εκτός από τις ντετερμινιστικές μεθόδους που εφαρμόσαμε, με έμφαση στην πιθανή δυνατότητα των τεχνικών αυτών, για ανάλυση της σημασίας του προς εκφώνηση κειμένου (semantics).

σημειώστε την λέξη που ταιριάζει περισσότερο στην λέξη που ακούσατε
[με τονισμένους χαρακτήρες είναι οι λέξεις που τελικά εκφωνήθηκαν]

1.	μόνος φόνος	κώνος γόνος	πόνος τόνος	20.	λίμα κύμα	ρήμα χύμα	νήμα βήμα
2.	καις δεσ	πες θες	λες μπες	21.	δίκη Βίκυ	νίκη Μίκυ	λύκοι οίκοι
3.	χάρη βάραι	κάρυ φάραι	ζάρη πάρει	22.	ρήξη λήξη	νύξη δείξει	πήξη βήξει
4.	κότα φώτα	νότα μπότα	ρότα γιώτα	23.	τομή δομή	βωμοί ζωμοί	νομή χωνί
5.	με ντε	ρε ναι	σε και	24.	ταίρι βέροι	χέρι γέροι	μέρη ξέροι
6.	πίτα βήτα	κοίτα σήτα	ρίτα θήτα	25.	μόνοι χώνει	τόνοι κώνη	φόντοι ζώντοι
7.	χώρα Νόρα	μπόρα Ντόρα	δώρα φώρα	26.	κάμπος κάβος	Κάσος κάπως	κάλλος κάδος
8.	πείνα μήνα	χήνα Νίνα	φίνα δείνα	27.	καλώ κακό	καρό καπό	κανό
9.	πάνω πάκο	πάσσο πάρω	πάγο πάλλω	28.	κάπα κάρα	κάσα κάψα	κάβα κάρμα
10.	όπως όρος	όμως όντως	όλος όνος	29.	κύτος κύβος	κήπος κύρος	κοίλος
11.	πόλος θόλος	βόλος μόλος	ρόλος δόλος	30.	φάπα κάπα	τάπα πάπα	μάπα
12.	πέρας γέρας	τέρας κέρας	δέρας ξέρας				
13.	ρόδα ρότα	ρόζα ρόκα	ρώγα ρόμπα				
14.	λύκος λίθος	λίπος λήρος	λίγος λείος				
15.	λιτός λειψός	λιμός λινός	λοιπός Λισσός				
16.	φάρος βάρος	θάρρος σάρος	χάρος Πάρος				
17.	πόρος πόλος	πόνος πόσος	πόθος Πόντος				
18.	μήτε μπείτε	δείτε πιείτε	πείτε ζειτε				
19.	χύνω λύνω	δίνω ξύνω	πίνω ντύνω				





Βιβλιογραφία

Ελληνικές πηγές

1. Βαλεοντής Κ. & Καγιάφας Π. , Συγκριτικά Φωνολογικά & Γραμματολογικά Στατιστικά Στοιχεία της Ελληνικής Γλώσσας, 5^ο ετήσιο Συνέδριο Ελληνικής Ακουστικής Εταιρείας, Αθήνα 1984.
2. Βαλεοντής Κ. (1981), Φωνηματική Στατιστική Ανάλυση της Ελληνικής Γλώσσας, Επιστημονικό Συμπόσιο “Ακουστική”-Βόλος, Ιούνιος ’81
3. Κολίτσης Α.Γ. Επισκόπηση Τεχνικών Ανάλυσης και Επεξεργασίας Σημάτων Φωνής, διπλωματική εργασία, τμήμα Ηλεκτρολόγων Μηχανικών, Θεσ/νίκη 1997
4. Πανάς Σ. Ψηφιακή Επεξεργασία Σημάτων, University Studio Press, Θεσ/νίκη 1997.
5. Παπανικολάου Γ. (1985), Ηλεκτρακουστική, University Studio Press, Θεσ/νίκη
6. Σουρλαντζής Ι. (1996), Αυτόματη Εκφώνηση Κειμένου με Η/Υ, για την Ελληνική Γλώσσα, μεταπτυχιακή εργασία, τμήμα Πληροφορικής Πανεπιστημίου Κρήτης
7. Τριανταφυλλίδης Μ. Νεοελληνική Γραμματική της δημοτικής, ανατύπωση της έκδοσης του Ο.Ε.Σ.Β. Θεσσαλονίκη (1996).
8. Τσολάκης Χ. & Χατζημιχαήλ Χαρούλα (1998), σημειώσεις για την Προσωδία του Λόγου.

Ξένες πηγές

9. Ainsworth W.A. Pell B. (1989), Connectionist Architectures for a Text-To-Speech System, Proceedings of Eurospeech '89, vol 1.
10. Allen J., Hunnicut S. & Klatt D. (1987), From Text To Speech, The MITALK System, Cambridge University Press.
11. The Bell Labs Approach. Multilingual Text-to-Speech Synthesis, Kluwer Academic Publishers, 1998.
12. Charpentier F. & Stella M.G. (1986), Diphone Synthesis using an overlap-add technique for speech waveforms concatenation, Proceedings of the international conference on acoustics, speech and signal processing '86
13. Charpentier F. & Moulines E. (1989), Pitch Synchronous waveform processing techniques for text-to-speech synthesis using diphones, Proceedings of Eurospeech '89, Paris vol. 2
14. Donovan R.E. (1996), Trainable Speech Synthesis, dissertation @ Cambridge University, Engineering Department.

15. Dutoit T. & Leich H. (1993), MBR-PSOLA: Text-to-Speech Synthesis based on a MBE resynthesis of the segments database, Speech Communication, n° 13
16. Dutoit T. (1997), An introduction to text-to-speech synthesis, Kluwer Academic Publishers
17. Hunnicut S. (1980), Grapheme-to-phoneme rules: a Review, Speech Transmission Laboratory, QPSR 2-3, Royal Institute of Technology, Stockholm, Sweden
18. Klatt D.H. (1980), Software for a Cascade/Parallel Formant Synthesizer, Journal of Acoustical Society of America, vol. 67, No 3
19. Klatt D.H. (1987), Review of Text-to-Speech Conversion for English, Journal of the Acoustical Society of America, Vol. 82, No 3
20. Logan J.S., Greene B.G. & Pisoni D.B. (1989), Segmental Intelligibility of Synthetic Speech Produced by Rule, Journal of the Acoustical Society of America, vol. 86, No 2.
21. Markel J.D. & Gray A.H. (1976), Linear Prediction of Speech, Springer-Verlag, Berlin
22. Oliveira L.C. (1993), Estimation of source parameters by frequency analysis, Eurospeech, Vol 1, Berlin
23. Oliveira L.C. (1994), Text-To-Speech Synthesis with Dynamic Control of Source Parameters
24. Rabiner L.R. & Schafer R.W, Digital processing of speech signals, Prentice-Hall, 1978
25. Rosenberg A.E, Effect of glottal pulse shape on the quality of natural vowels, J.A.S.A. 49,(2, part 2), 1971
26. Van Coile B. (1991), Inductive Learning of Pronunciation Rules with the Depes System, Proceedings of the International Conference on Acoustics, Speech & Signal Processing '91, Toronto, vol 2.

Επιλεγμένες Πηγές από το Διαδίκτυο

<http://asa.aip.org/>

<http://www-dsp.rice.edu/software/>

<http://sig.enst.fr/~speech/index.html>

<http://www.esca-speech.org/>

<http://tcts.fpms.ac.be/synthesis/mbrola.html>

<http://tcts.fpms.ac.be/synthesis/modelcmp.html>

<http://www.cstr.ed.ac.uk/projects/festival.html>

<http://www.dvsinc.com/iambe.htm>

<http://www-dsp.rice.edu/software/RU-FILTER/>

http://www.media-enterprise.de/home_eng.htm
<http://www.cse.ogi.edu/CSLU/>
<http://www.phon.ucl.ac.uk/resource/cal-top.htm>
<http://leb.net/blinux/>
<http://www.research.microsoft.com/msrinfo/demodwnf.htm>
<http://www.ims.uni-stuttgart.de/phonetik/gregor/synthspeech/>
<http://www.cs.princeton.edu/~prc/>
<http://sig.enst.fr/~pierrot/SigLabAdress.html>
<http://speech.essex.ac.uk/speech/material/mark/111/lg111-2b.html>
<http://macserver.haskins.yale.edu/Haskins/HEADS/MMSP/intro.html>
<http://www.ling.su.se/staff/hartmut/kemplne.htm>
<http://mambo.ucsc.edu/psl/speech.html>
<http://xxx.lanl.gov/cmp-lg/>